A Computational Inquiry Into Navigation With Particular Reference To The Hippocampus

D. J. Foster

Submitted to the University of Edinburgh

for the Degree of Doctor of Philosophy

September, 1999

In accordance with the University of Edinburgh postgraduate study regulation 3.8.7, I declare that this thesis was composed by myself, and the work presented herein is my own.

D. J. Foster

To Mum, Dad, Joanna and Simon

Acknowledgements

I would first like to thank my supervisors. As principal supervisor, Richard Morris is due thanks for giving me the initial opportunity of studying for this degree. Since then, he has provided a mixture of solid advice and inspiration – especially in lab meetings. As for my second supervisor, Peter Dayan, it has simply been a privilege to work with him. None of this thesis would exist without him, nor perhaps my desire to continue on a related course.

I have been extremely lucky in the people I have been able to work with. The students and ex-students in Edinburgh, especially Livia de Hoz, Steve Martin, and Mark Ramsay, made work and play a pleasure, and whatever happens next I will miss them. I should also like to thank Gill Maitland for help on numerous occasions. At M.I.T., there are too many people to mention, except Prof. Matthew Wilson, Andrea D'Avella, and the residents of The Hub. Outside of work, without Drs. J and M Hutchison, none of this would have been possible; Mum, Dad, Joanna and Simon, likewise; and a few others made all this turn out the way it has, among them Andrew Salway, Peter Roper, Declan Colgan and Pauline McEwan. I wish, as well, that there were a way I could thank Keith Ruddock, whose kindness to me many years ago meant so much.

I should also like to thank the Faculty of Medicine, Edinburgh University, for a Holdsworth scholarship; the Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, U.S.A., for support during my visits there; the McDonnell-Pew Centre for Cognitive Neuroscience, Oxford, for support with travel costs; and the Gatsby Charitable Foundation for support during completion of part of the work described in this thesis.

Abstract

Goal-directed navigation is the art of traversing space in order to get to a goal. Animals can be expected to learn the capacity to navigate optimally, that is, learn to be able to take direct paths from where they are to where they want to go. However, optimal local behaviours are determined by global constraints. A common theoretical suggestion is that in order to navigate animals build and use a global representation in the form of a map. Maps pose two difficult problems, however: a global-from-local problem in building a consistent map, and, of greater difficulty, a local-from-global problem in reading the map. An alternative to using a map is to learn appropriate local behaviours directly from interacting with the world, using only local representations. This thesis reviews evidence that a neural structure in the mammalian brain known as the hippocampus provides a representation of space that is purely local. The thesis then reviews a selection of computationally efficient, neurally plausible, reinforcement learning methods that can use local representations to learn optimal actions in navigation-like tasks. The key disadvantages usually associated with the methods are slow learning, and inflexibility to change. Both disadvantages were investigated in this thesis in the context of learning to navigate.

A first model is presented which learns to perform two hippocampally dependent tasks, one involving navigation to a single goal, the other involving navigation to multiple goals. Animals gradually acquire the single goal task, but in the multiple goals task, gradually acquire the ability to navigate directly to a novel goal on only the second trial to that goal. One component of the model uses place cells within a standard reinforcement learning scheme, temporal difference learning in an actor-critic. This component by itself captures performance in the single goal task, but fails to capture one-trial learning in the multiple goals task. A second component of the model learns globally consistent coordinates from local self-motion information, in a novel application of temporal difference learning. This coordinate learning is relatively independent of the behaviour of the animal, enabling the gradual acquisition of one-trial learning in the multiple goals task to be captured. Two purely behavioural predictions follow which were tested experimentally. First, once a coordinate system has been learned, simple placement of an animal at a novel goal should provide the animal with sufficient information to allow direct paths on the next trial, and this was shown to be true. Second, since coordinates offer no principled way of circumnavigating barriers, the model predicts that animals will be unable to learn the task in the presence of barriers, but this was shown to be false.

A second model is presented, based on the idea that *unsupervised learning* can be used to find structure in value functions (the predictions of reward learned by reinforcement learning methods), that corresponds to structure in the environment, *eg* due to the presence of barriers. Decompositions of various environments are found of different types – flat, linear and hierarchical. Results show that that structure learned from just a few goals compares well with that learned from all possible goals, and that augmenting a standard state representation with these decompositions leads to faster acquisition, again within the purview of the local, neurally plausible reinforcement learning methods. These results emphasise the importance of representational learning, and suggest that with appropriate representations, simple learning methods can rival more rigidly specified and complex methods.

Contents

Chapter 1

Introduction

Navigation, as I wish to consider it, is the art of traversing space in order to get to a goal. A facility with navigation is surely fundamental to any animal that needs to return to goals in its environment, such as food, water, nest or burrow. The aim of this thesis is to describe a computational inquiry into this facility, focusing on the one hand on the kinds of computations that animals may need to carry out in order to navigate, and on the other hand on the particular contribution of a region of the mammalian brain known as the hippocampus. The subordinate aims of this introduction are therefore first to clarify the rationale, motivation and scope of the approach, and second, to present a description of navigation in terms of its inherent computational problems.

1.1 Rationale, Motivation and Scope

The idea that to understand neural processes one first needs a characterisation of the computational problem exercising those processes was proposed by Marr (1982). Three features in particular were identified as necessary:

1. **A problem-based approach.** Marr suggested that in order to understand how neural systems were solving some problem, it was necessary to be able to describe how the problem might be solved at all. The key insight is that most areas of neuroscience are concerned with problems that human engineers cannot solve. Often this goes unnoticed: "In the 1960s almost no one realized that machine vision was difficult...The reason for this misperception is that we humans are ouselves so good at vision...the idea that extracting edges and lines from images might be at all difficult simply did not occur to those who had not tried to do it". Navigation presents just as many pitfalls as vision, whose artful solutions are kept just as hidden.

2. **A focus on underlying principles.** Marr noted that solutions to problems of an *ad hoc* nature were less likely to provide insights into the nature of neural processing than a knowledge of the underlying principles governing solutions. Marr's example was trying to understand the function of feathers: an understanding of aerodynamics is required, yet the principles of aerodynamics underly *every* earthly flying machine. One would be suspicious of an alternative theory that did not generalise in this way.

3. **Generality of perspective**. Marr also identified a rather more subtle requirement for the computational approach, that it would not suffice to solve overly simplistic versions of problems in the hope that the solutions would, in some sense, scale up to the more realistic problems. In nascent artificial intelligence, to use Marr's example, solutions to deterministic problems often failed to provide even appropriate intuitions for solving related but more realistic stochastic problems.

The approach taken in this thesis is to proceed from just such a computational level of analysis for navigation. In the following sections of this introduction, a series of questions is asked. First, how should navigation be defined? Second, what inherent computational problems arise given this definition, and what *algorithms* exist that might solve these problems? The central issues are what representations might be learned, and what processes might make use of them.

It may be remarked that the algorithms the brain uses to navigate will not have been shaped solely by the computational problem which the algorithms must address, but also by biological constraints. Indeed, the point was fully acknowledged by Marr: "the algorithm... also depends on the characteristics of the hardware in which it is to be implemented. For instance, biological hardware might support parallel algorithms more readily than serial ones, whereas the reverse is probably true of today's digital electronic technology". Marr supported the notion of neural constraints, but an important insight was that these constraints are weak. Generally speaking, one can rule out algorithms on the basis of neural arguments, but it is hard to *specify* algorithms from neural constraints. One reason why this is so is because of the computational power of networks of neurons – even quite simple neural networks are capable of transforming input information in virtually unlimited ways.

Neural constraints are incorporated into the algorithms investigated in this thesis by the following means:

4. **An attention to behavioural neuroscience data**. Behavioural neuroscience studies have been invaluable in discerning the separable roles of different areas of the brain in various tasks. Lesion, pharmacological and molecular-genetic studies have been particularly valuable in identifying the *necessary* roles played by certain brain areas

during various tasks. Electrophysiological studies have provided compelling characterisations of the *representations* being used in certain brain areas – a key component of Marr's algorithmic level of description.

5. **A connectionist framework**. Connectionist, or neural network, models are extremely simple characterisations of neurons and the way they represent and compute information. Although simple, they can constrain the algorithmic level in some basic ways. For example, connectionist computations tend to place an emphasis on the *representation* of information, in particular on multiple, distributed representations (in terms of multiple layers of neurons that are active in parallel); a separate emphasis is placed on *learning* (in terms of changes in the strengths of connections between neurons); an overall emphasis is placed on the *statistical* nature of the computations; and computations themselves tend to be *iterative*, *local* to active neurons and the connections between them, and *online* during the receipt of information or the production of behaviour.

The main inspiration behind this thesis, however, is the recognition that a large body of relevant work has been generated by a field in artificial intelligence whose sole concern is to solve (at both computational and algorithmic levels) difficult problems of a kind very similar to navigation (Watkins, 1989). This field, reinforcement learning (RL), is problem-based to the extent that in both theory and practice RL methods are judged in terms of solving difficult problems, is principled in that much of its theory is based on a fundamentally mathematical characterisation of the problem, and is, at least in comparison to many artificial intelligence approaches, general in its perspective, in that issues such as stochasticity are axiomatic. Moreover, many RL methods have a natural expression in terms of connectionist networks, and there is even evidence for neural processes performing the roles of components of basic RL algorithms.

RL is an approach to problems in which rewards and punishments predominate, that is to say, problems in which the information the environment provides to an animal, and which it can use in order to learn things, is extremely sparse, both in terms of when and where it is available (*eg* only at a food source) and what it consists of (*eg* simply the existence or non-existence of reward). However, this sparsity should be distinguished at the outset from "algorithmic sparsity", which might be said to underly theoretical approaches to animal learning such as stimulus-response (S-R) learning (*eg* Hull, 1943). Indeed, supposing that constraints at the level of the problem must translate into somehow equivalent constraints at the level of the algorithm or of neurons is what Marr would have called a confusion of levels. RL is defined by a problem statement, that is, a computational level of analysis. This places a key constraint on putative algorithms in the form of dealing with the identified

problems effectively, which can demand richly structured algorithms. By contrast, too simple algorithms such as S-R learning fail to satisfy this constraint.

There is one further aspect of RL which motivates its consideration. It offers an explanation for a paradoxical phenomenon in the neuroscience literature, with which this thesis will be extensively occupied: the role of hippocampal neurons in mammalian navigational learning. The hippocampus is a structure in the mammalian brain that is likely to be necessary for learning to navigate, but the activities of neurons of this structure, while intriguingly suggestive of a role in navigation, nevertheless are limited in ways which make their contribution to navigation difficult to understand. Models derived from RL offer one possible answer.

The scope of this inquiry is limited in various ways. First, due to the preoccupation with the hippocampus, the focus is necessarily on mammalian navigation. A large literature concerning navigation in other classes of species, such as insects and birds, will not be discussed in much detail. Second, within this class, there is a natural focus on the behaviour and neural functioning of the rat. This is arguably desirable not just because of the availablity of experimental data, but also because, it is hoped, the range of behaviours being narrower and the influence of phylogenetically newer neural areas such as prefrontal cortex being smaller, the rat presents a more transparent system in which to study hippocampal function. There are arguments against this position, however, and so an attempt will be made to relate conclusions to what is known about the hippocampus in other mammals. Third, various aspects of navigation which can be attributed to various other computational systems will indeed be so. For example, a variety of difficult computational problems can be anticipated in the task of generating movements such as walking or running, but these kinds of problems will not be looked at in this thesis. Instead, it will be assumed that provided actions can be specified appropriately, other neural systems will execute them. Likewise, problems in vision, audition and somatosensory sensation will not be investigated. The following sections, which map out a level of description for navigation, will hopefully make these omissions clearer.

1.2 Navigation Is An Optimisation Problem

Navigation might be defined simply as getting from where you are to where you want to be. This is too simple, however. An aeroplane can fly from Edinburgh to London via Australia but presumably the route would be unacceptable to most passengers. Random choices of actions can, in many situations, be guaranteed to take an animal from one place to another, but are unlikely to be adaptive. The simplest meaningful definition of navigation is getting

from where you are to where you want to be as quickly as possible. Alternatively, one might say, by the most direct route possible.

There is clear selection pressure for the ability to navigate efficiently. For example, a prey species such as the rat will survive longer the more direct its paths are to its burrow when a predator threatens. What is optimal, however, depends greatly upon the task in hand. The behaviour of a satiated animal, for example, would be poor evidence against the hypothesis that the same animal could take a direct path to a remembered food source if it wanted to. The suggestion is that navigation is a *critical skill*, in Watkins' (1989) sense: it serves animals well to learn how to navigate efficiently, for just those times when such a capacity is needed.

The idea of navigation as an optimal process can appear remote from our everyday experience. However, optimality assumptions underpin many behavioural experiments. For example, a common measure in behavioural experiments is latency, the time taken by an animal to reach a goal. Short latencies are offered as evidence that the animal knows how to get to the goal, and long latencies as evidence that it doesn't. In fact, every latency may have been followed by attainment of the goal – but we cannot credit that an animal would know how to get to the goal and still choose a very inefficient path. Naturally, our interpretation of the results reflects a belief that the experimental design incorporates sufficient motivation for the animals to reveal their navigational capacities. Behavioural results from navigational tasks are used throughout this thesis, to motivate models of navigational learning, and to benchmark the performance of the models in simulated versions of the tasks.

What if not all routes are equal? Consider this common choice: route A might be shorter but usually involves heavy traffic, while route B is fast but there is an expensive toll to pay. To define an optimal choice, a comparison must be made between the cost of sitting in traffic and the cost of paying the toll. Although it would appear difficult to compare the two, nevertheless anyone who makes a non-random decision *is* making the comparison. If one can assign costs to seemingly different sorts of event, then a general definition of optimal can be utilised: an optimal path is a path that incurs the minimum total cost, out of all the possible paths that might have been taken. This can be generalised to include positive reward information: optimal paths maximise rewards and/or minimise costs. Navigational behaviour has a natural definition in terms of reinforcement.

Note that putative costs and rewards are really only a surrogate for the "true" costs and rewards in the environment, and the extent to which optimal behaviour with respect to the former is valuable depends on their relation to the latter. For example, costs associated with excessive detours on a rat's return path to its burrow are useful because they reflect among

other things the potentially mortal cost of such detours if a predator is nearby. It is assumed here that animals are informed by appropriate costs and rewards.

1.3 Computational Problems Inherent In Navigation

The decision facing a navigating animal at any moment is to choose from a set of *local* behaviours, such as moving forwards, or along a path, or towards a visible object. While pondering its decision, the animal perceives a *local* world of landmarks, other objects and perhaps a sense from its own movements of what it is itself at that moment doing. There may even be *local* costs or rewards available, dependent on the behaviour taken and the part of the world the animal is in.

Unfortunately for the animal, these three aspects of its experience cannot be related in isolation. The correct behaviour locally depends not just on the local environment but also on where the distant goal is, and on all the rewards or costs that might be incurred while navigating to it. Navigation is a *global* problem.

1.3.1 The Global-From-Local Problem

An intuitive proposal for dealing with the global aspect of navigation is to learn a global representation of the world. For example, the most widely accepted theory of how animals navigate holds that they build and use an internal representation of the world in the form of a *cognitive map* (Tolman, 1948; O'Keefe and Nadel, 1978; Gallistel, 1990). Chapter 2 of this thesis discusses some of the neural issues raised by the claim that a structure in the mammalian brain known as the hippocampus instantiates this map (O'Keefe and Nadel, 1978; McNaughton *et al*, 1996), drawing attention to the apparently *local* nature of the representation provided by the neurons of that structure. The discussion here focuses on the computational implications of the proposal.

The first issue for a global representation is how it may be constructed. McNaughton *et al* (1996) imagine a preconfigured but empty map into which animals place object information as they move around. This resembles the earlier conception of O'Keefe and Nadel (1978): "the animal brings to each new situation a *tabula rasa* of potential place representations. One is chosen to represent a specific location... [which] *automatically determines the way in which the remainder of locations will be represented*" (emphasis added).

However, putting information into a global map means knowing first where on the map to place the information, which in turn implies "self-localisation" (Redish and Touretzky, 1995), that is, knowing the global coordinates of ones location at every moment. While

some measure of coordinate-like information might be obtained from various sources, such as from the way the view of an array of landmarks changes as one moves past it, or from ones own sense of movement integrated over time, these sources of information are fundamentally local. Coordinates from different local sources will be globally inconsistent.

This global-from-local problem is well recognised, and many schemes for building global representations from local patches of experience have been suggested (Cartwright and Collett, 1987; Worden, 1992; Wan *et al*, 1993; Foster *et al*, 1999, and chapter 4). However, the question arises of whether the effort involved is worth it.

1.3.2 The Local-From-Global Problem

Consider the ways in which one might read a map in order to navigate. One way is to draw a straight line from the current location to the location of a goal. There are advantages to this scheme: it is fast and easy, and it is flexible, specifying a bearing for an arbitrary choice of goal. However, two points may be made about this usage: (1) Finding a global direction is not a general solution to the navigation problem. For example, from my desk writing this I can point directly to my favourite sandwich shop just outside the building, but the direction in which I point has no relation whatsoever to any of the directions in which I would have to move, at any point within the building, to get there. Only in rather special circumstances might a global direction prove useful. (2) A corollary of this result is that a whole map of information is rather redundant for the purpose of just working out the global bearing of a goal – only the current coordinates and the remembered coordinates of the goal are required, as no use is made in the calculation of the coordinates of any other object. Therefore, the view is taken in this thesis that while global direction may in certain circumstances be useful – and one such circumstance is modeled in chapter 4 – nevertheless global direction should be a *specialised* mode of control, requiring less learning than for a whole map, and offering limited and optional advice.

A second way to read a map is, I would argue, the way most of us do in fact read maps. It requires applying a range of complex visual search routines in order to compute an appropriate route from current position to goal. The information we need is certainly contained in the map, but it is nevertheless quite a feat to turn this information into a sensible choice of actions. The source of the difficulty is a combinatorial explosion in the number of paths to be considered. To get to my sandwich shop I pass through about eight places at which there is a choice of at least three different routes. If all such choice-points in the world offered only three choices, and I had no prior expectation about which choices to make, a map would force me to search through at least $3^8 = 6561$ paths to find out which were shortest – and a good proportion of these to find one that worked at all. This is an unac-

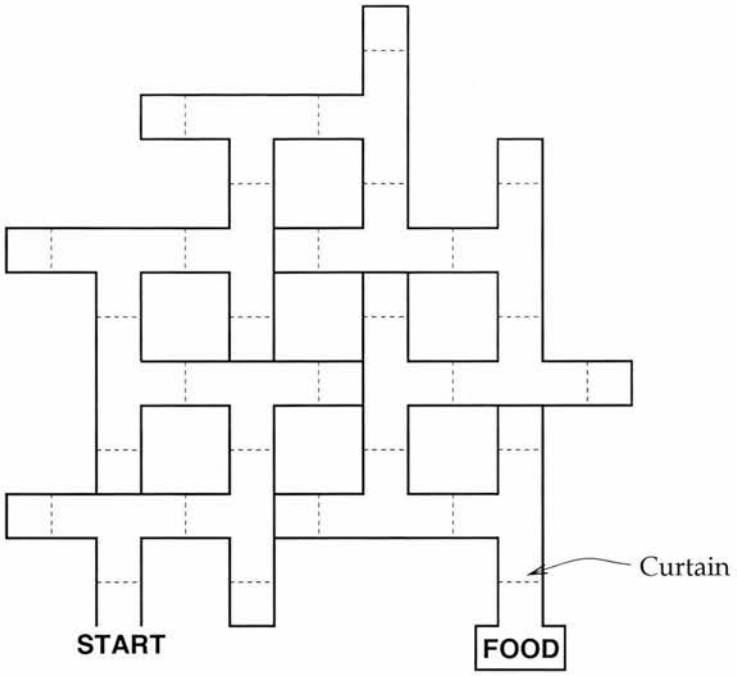


Figure 1.1: The maze used by various groups investigating latent learning in the 1920s and 1930s. At each choice point, the action which does not lead *immediately* to a dead-end is the ultimately correct one. After Tolman (1949).

ceptably inefficient mode of planning. A second example might be a taxi driver who, with meter ticking, planned routes around a city, nose stuck in a map. One would not accept this kind of behaviour because a taxi driver is paid for, among other things, familiarity with his or her environment. The possibility which we don't consider for the taxi driver, but which cognitive map theory would have us consider for an animal, is that he or she might be perfectly familiar with an environment and still choose to navigate using a map.

The local-from-global problem, then, is that local behaviours need to be specified, but are not made explicit by a global representation such as a map. In rather special circumstances, a globally calculated direction might be useful, but this is not generally the case.

1.4 Solving The Navigation Problem Locally

The problems identified in the previous section can be illustrated with an example. One of the best known results from the very early history of experimental psychology was that of "latent learning", in which rats were thought to learn about the spatial structure of their maze environment, and rapidly – in a single trial – turn this knowledge into appropriate actions upon discovery of a reward (Blodgett, 1929; Elliott, 1929; Tolman and Honzik, 1930b). Indeed, O'Keefe and Nadel (1978) cite the early latent learning experiments as a key influence on their theory of the cognitive map. In the basic experiment, two groups of

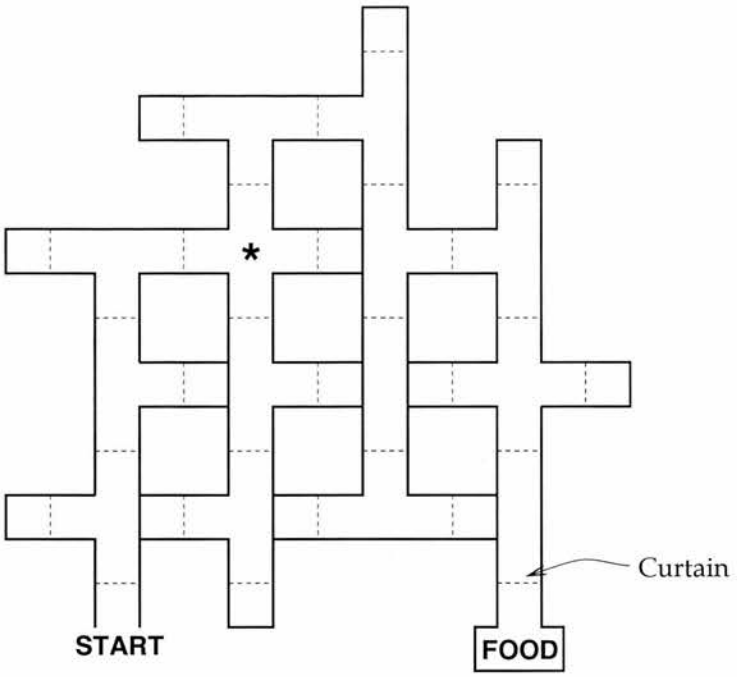


Figure 1.2: A more challenging latent learning maze, in which the ultimate appropriateness of choices is not *locally* specified, but has to discovered from *global* information. For example, paths from one choice point (*) do not immediately reveal whether they are appropriate or not, but must be searched through exhaustively.

rats were trained on a complex maze. Rats in the first group were rewarded with food at the goal throughout training. Rats in the second group were unrewarded for a number of trials, and subsequently rewarded on all further trials. The pattern of results was in each case striking – previously unrewarded rats improved dramatically the very next trial after being rewarded, in terms of both errors made, and time to get from start to finish, and in fact caught up with the performance of the rats that had been rewarded all along. Because, it was argued, the unrewarded rats were unable to learn goal-specific responses gradually like the rewarded rats might have, therefore the one-trial learning of appropriate responses implied the use of a global representation of the environment, such as a cognitive map (Tolman 1948; O'Keefe and Nadel, 1978).

The problem with the result, however, is the maze itself ("fourteen-unit T-maze", figure 1.1; Blodgett, 1929, used a "six-unit T-maze"), which is so constructed as to remove the *global* problem completely! A single-step look-ahead suffices to tell the rat which is the ultimately correct choice. While curtains obscure a perceptual solution to the problem, the demands in terms of learning remain fairly trivial, in effect learning a response for each choice point on the basis of an immediate dead-end.

However, simply by removing or rearranging internal walls in figure 1.1, a rather more challenging "latent learning" experiment can be constructed, as shown in figure 1.2. The

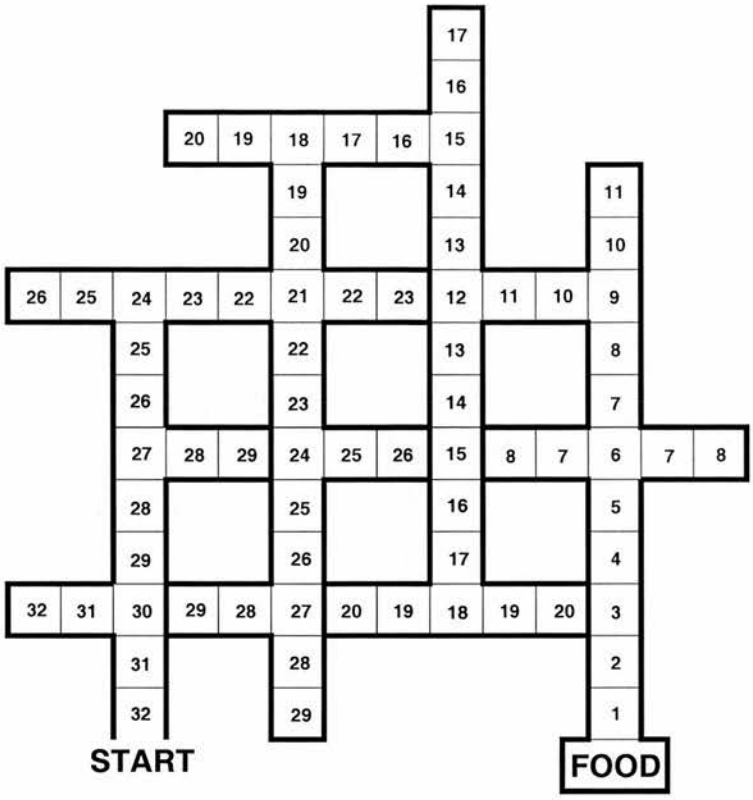


Figure 1.3: The value function for the maze problem of figure 1.2. Each location has an estimate of minimum total future cost associated with it, assuming costs of 1 for every move not onto the goal.

difficulty with this maze, which is absent from the previous one, is that the correct choice is determined globally rather than locally. Given a map of the maze, such as figure 1.2 itself, there is a search problem as described in the previous section.

How might this task be solved efficiently? One answer, that has its roots in a set of engineering methods known as dynamic programming, is to develop for each location an estimate of the minimum total *future* cost that might accrue from the moment of occupying the location to the moment of reaching the goal. Under the assumption that a cost of 1 is always received after any move not onto the goal, the complete set of estimates for all locations, known as the *optimal value function*, would have the form indicated in figure 1.3. The key observation is that, as a representation, the value function all but solves the local-from-global problem in navigation: by choosing at each step a behaviour that minimises the cost of the next state, an optimal path may be recovered. Expensive search strategies are not required.

As alluded to in the opening section, a key motivation behind the work in this thesis is the fact that recently developed reinforcement learning algorithms are able to learn an optimal value function in a purely online and *local* manner – by refining local estimates of

value while attempting to navigate. A selection of these methods is reviewed in chapter 3. The only representational requirement for these methods is a representation of state that individuates locations. A global representation is not required.

Two main drawbacks are associated with reinforcement learning methods. The first drawback is that in many simulated tasks (which cover a wider range than just navigation tasks) they can be slow to learn. Furthermore, performance scales poorly with the number of states. The second drawback is inflexibility to change, such as a change of goal position, or local changes to the structure of the environment. The next two sections consider these two issues in turn, and suggest, perhaps counter-intuitively, that an appropriate *local* representation can be effective in addressing each.

1.5 Modeling Behavioural Data

One constraint to impose on a model of navigational learning is that it should be able to account for actual data from animals in a navigational learning task. This is particular important given the speed of learning issue, identified as a possible drawback for the application of reinforcement learning methods to navigation.

In choosing appropriate data to model, two concerns were paramount. First, the data were to come from a well understood and modern spatial task, in which there is adequate control for trivial navigation strategies such as heading towards a perceptible local cue, such as an odour or visible reward site. Second, because a key component of the model to be investigated is the representation of space provided by neurons of the hippocampus, the spatial tasks were required to be hippocampally dependent. The evidence for the hippocampal dependence of the tasks described in this section is reviewed in chapter 2.

A popular task for investigating spatial learning, and the hippocampal contribution to it, is the radial arm maze (Olton and Samuelson, 1976). The apparatus consists of a central platform with a number (usually 8 or greater) of equally radially spaced arms leading outward from it. Each arm usually contains no local cues distinguishing it from any other; rats must use the rich array of distal cues in the environment to distinguish between arms. The basic experimental technique is to bait the arms, *ie* place food at the end of an arm, an important assumption being that the rat cannot ascertain the presence or absence of food perceptually from the central platform. A hungry rat is then placed on the central platform and is expected to consume the baits one by one, and with few *working memory* errors – defined as re-rentries into an already visited (*ie* exhausted) arm. A common concurrent procedure is consistently across trials not to bait certain arms – so allowing *reference memory* errors, defined as entries into these never-rewarded arms. Normal rats acquire the task, making on

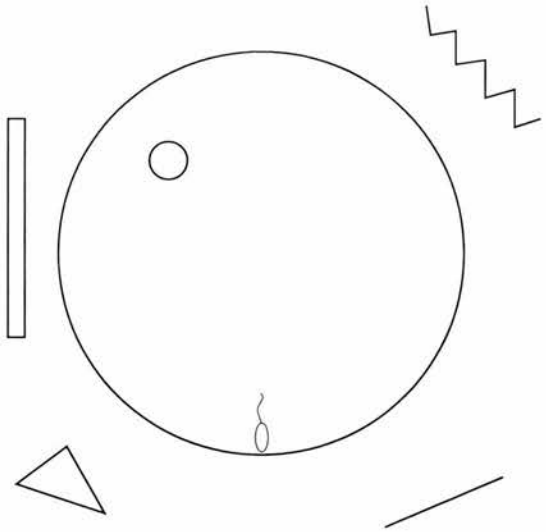


Figure 1.4: The watermaze apparatus: rats are placed in at the edge of a pool and are motivated to swim to a platform which they cannot see directly. Instead, distal cues in the surrounding room must be used to support navigation.

average around one working memory out of the first eight arm choices (Olton and Samuelson, 1976), and are also able to learn to avoid making reference memory errors (Olton *et al*, 1979).

However, there are a number of difficulties with this task as a suitable candidate for modeling navigational learning. First, an action choice need only be made from one, locally-defined location, before a reward is received, because of the trivial nature of following an arm to its end. So the navigation problem is completely local, and the only difficulty in defining location comes in maintaining an allocentric sense of orientation. However, even this might be suspect for the working memory component of the task, because it is only necessary that an animal maintains a sense of orientation that is consistent during a single trial, which could in theory be supported by dead-reckoning its orientation from a short-term record of its own movements. Thus, Brown and Bing (1997) report excellent performance by rats using a radial maze with a central platform from which extramaze cues could not be visually perceived. Therefore, the radial arm maze might be an appropriate way of studying certain kinds of memory, but is not necessarily an appropriate task with which to study navigational learning.

A second modern navigation task which limits the usefulness of local cues rather more severely is the reference memory in the watermaze (RMW) task (Morris, 1981; Morris, 1984). The watermaze is a circular tank of water (diameter 2m), the walls of which are painted a single colour (figure 1.4). At some location within the tank, a solid platform (diameter 11cm) rests just below the surface of the water, rendered invisible to the swimming rats by refraction through the water which is made slightly opaque by the addition of

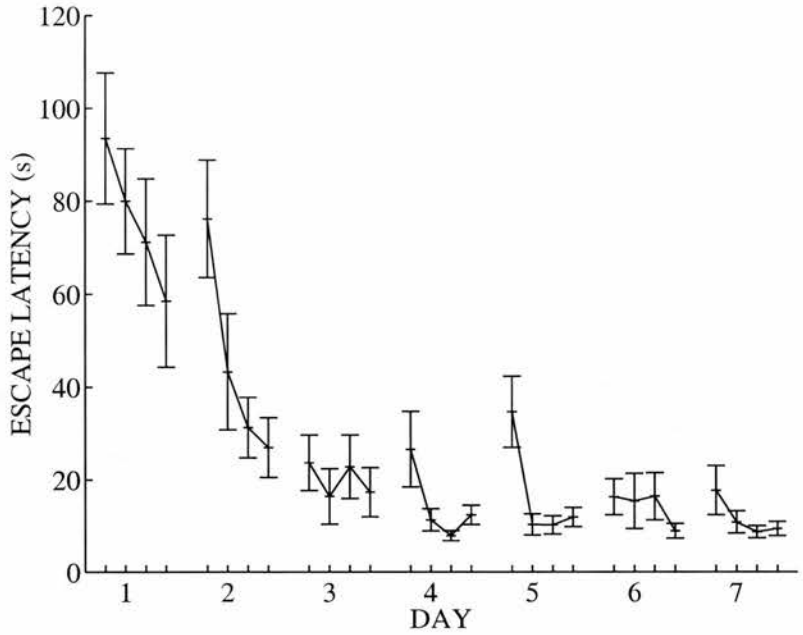


Figure 1.5: The performance of rats on reference memory (RMW), N=12. Mean escape latency (time taken to reach the platform) is plotted across days (4 trials/day, fixed platform location). Amended from Foster *et al* (1999).

powdered milk or latex. While local cues are limited, the apparatus is typically situated in a room rich with distal cues such as posters, curtains and metal stands. On each of several trials per day, a rat is placed in at the edge of the pool. To make the distal cues even more important, the starting position is varied in a random fashion from trial to trial between four equidistant positions. Clearly this also causes problems for simple dead reckoning strategies such as simply integrating over self-motion estimates.

The task reveals *optimisation* of navigational behaviour. Rats are excellent swimmers, and are also highly motivated to escape. However, the level of motivation appears not to be so great that rats are unable to explore sufficiently. Swimming at on average approximately 25cm/s, they take more or less direct paths to the platform after at most 20 trials, as implied by their short escape latencies (figure 1.5). During subsequent probe trials (or "transfer tests") with the platform removed, rats search mainly within the quadrant of the pool in which the platform was located (Morris, 1981).

Learning in the RMW task might be modeled by building a map – because the structure of the environment is so simple that a map might be read easily. A more minimal strategy, however, is to model RMW in terms of learning to predict rewards, using reinforcement learning methods. A critical aim behind this strategy is to determine whether these methods, in tandem with a hippocampal representation of space, can learn optimal paths as quickly as rats can.

1.6 Flexibility In Navigational Learning

An important aspect of navigation as a problem is the possibility of change. For example, the position of the goal might change, *eg* if a food source is exhausted or a burrow attacked, and newly established goals need to be returned to. As a different example, an aspect of the environmental structure might change, *eg* if a river crossing floods, or a well-used path becomes blocked in some way. The critical issue is the speed with which an animal can change its behaviour appropriately to the new situation; the adaptive significance of such fast relearning is clear. Addressing change is important because it is commonly perceived as a weakness of reinforcement learning methods. Moreover, exactly because a map is unable to specify actions locally but instead represents information globally, it copes after a fashion with both arbitrary goals and local changes to environmental structure. Such flexibility was highlighted early on as a key advantage for using a map (O'Keefe and Nadel, 1978).

Computationally, the key issue is generalisation. When the world changes in some way, are there aspects of the world that stay the same – or change in ways that are at least predictable? Can this underlying predictability be learned? If an animal can bring previous knowledge to bear upon a novel problem, acquisition of the problem should be faster than without the previous knowledge.

An important tool for developing models of this kind of generalisation is, again, actual behavioural data with which to compare performance. An interesting task might be the proposed "latent learning" experiment using the maze in figure 1.2. However, results using complex mazes are simply unavailable. While the history of experimental psychology of the first half of this century is rich in maze tasks, the results are invariably difficult to interpret because of the lack of appropriate cue control, or control for trivial navigational strategies.

1.6.1 Reinforcement Learning And Coordinates

Results are available, however, from a modern task that poses a multiple-goals problem. The task, developed by Morris (1983), Whishaw (1985, 1991) and Steele and Morris (1999), and following the last authors referred to here as delayed match-to-place (DMP), uses the same watermaze apparatus as the RMW task described above. Note that a watermaze appears the same wherever the goal is located. Several training trials per day are given to a platform that stays in the same location throughout the day. Critically, however, the platform is moved to a novel location before the start of each day. The latencies of figure 1.6 show that during the first few days, performance undergoes a gradual improvement

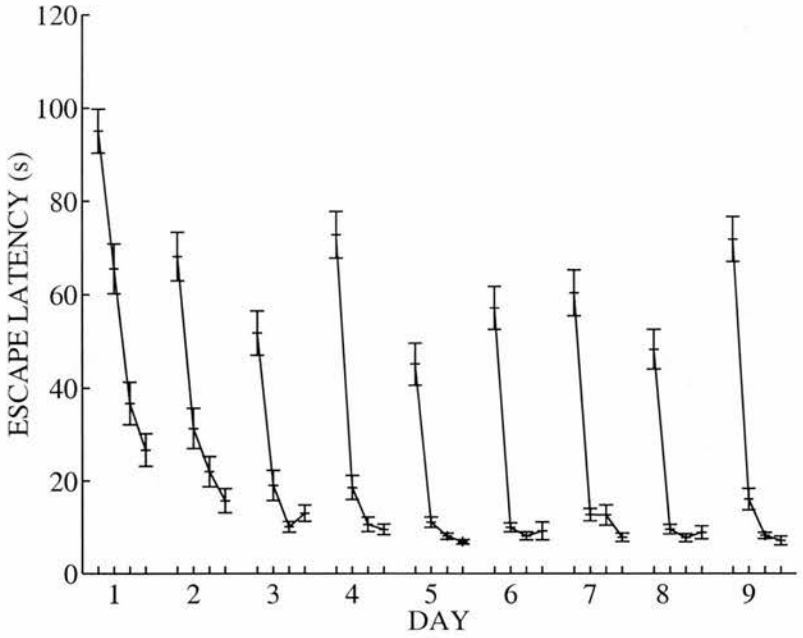


Figure 1.6: The performance of rats on delayed matching-to-place (DMP), N=62. Mean escape latency (time taken to reach the platform) is plotted across days (4 trials/day, new platform location each day). The pattern of latencies changes with day from a gradual improvement across trials on day 1 to one-trial learning by day 6. From Steele and Morris (1999).

within each day in the time taken to reach the platform (escape latency). A different pattern emerges by about the sixth day. Rats by then are showing *"one trial learning"* – that is, near asymptotic navigational performance on the second trial of the day to a novel platform position.

The watermaze environment is in fact one in which a simple, global representation of the environment in the form of a *coordinate system* might be useful in supporting navigation to arbitrary goals. As noted in section 1.3, however, the development of globally consistent coordinates poses a difficult learning problem. For example, one candidate local source of information would be internal estimates of self-motion, that are thought to be available to animals moving around an environment. Critically, however, the removal of an animal at the end of each watermaze trial, and its placement in at the start of the next trial at an unpredictable location, make simple strategies such as simply integrating over self-motion estimates ineffective. A novel approach to this problem is taken in this thesis (chapter 4), which presents a model of coordinate learning from self motion information in the DMP task, using the local (but environmentally stable) hippocampal representation of space to anchor the coordinates, and using a reinforcement learning algorithm, temporal difference learning, to learn them. A rather important property of this scheme is that coordinates can be learned independently of the behaviour of the animal – an important feature for tackling the DMP task. A second property is that the coordinate control is made to work within

the simple action selection regime, with the effect that not only is a separate switching mechanism not required, but coordinates automatically only come to be relied upon to the extent that they are useful, *ie* they can be ignored in environments for which they would not be useful.

1.6.2 The Underlying Structure Of Reinforcement Learning Problems

Clearly, however, coordinate learning cannot be expected to work in general environments, such as the complex maze of figure 1.2. A "latent learning" experiment might reward two sets of latent learners in *different locations*, most effectively locations demanding different choices at an earlier choice-point. While such data is unavailable (although chapter 5 presents one attempt at obtaining data from a one-trial learning watermaze task with barriers), nevertheless the general problem of fast relearning of similar tasks in complex environments seems a realistic navigation problem. Were navigation methods based on reinforcement learning to prove wholly unable to deal with this problem, it would be hard to justify considering them as models of navigational learning. In fact, a recent focus in reinforcement learning has been obtaining just such generalisation in complex environments (Singh, 1992; Dayan, 1995; Thrun and Schwartz, 1995; Dietterich, 1998; Hauskrecht *et al*, 1998; Precup and Sutton, 1998; Moore, Baird and Kaelbling, 1998).

It has already been argued that a key issue in reinforcement learning is how information is represented. In particular, the representation of current state (*ie* location), while assumed to be local, nevertheless may take many different local forms. For example, the representation provided by neurons in the hippocampus is a critical component in applying reinforcement learning methods to account for data from the RMW and DMP tasks.

As a second example, consider the maze of figure 1.2 and its associated optimal value function (relevant to the goal shown; figure 1.3). Reinforcement learning methods scale poorly with the number of states, and without the help of a coordinate system, learning about a new goal position would require relearning the value of all those states from scratch. However, the *representation* of state in this maze ignores aspects of the underlying structure of the maze. For a large number of states in and around the left-most vertical arm, for example, the navigational task would be the same for a large number of different goal positions on the more rightward side of the maze. This motivates making use of large local representations such as fragments (*eg* figure 1.7).

This thesis explores multiple goals problems in complex environments by attempting to learn *local* representations of state but at many different spatial scales. However, by starting from information only at the most local scale, the problem is how to determine larger scale representations. For example, figure 1.7 indicates a possible set of such representations for

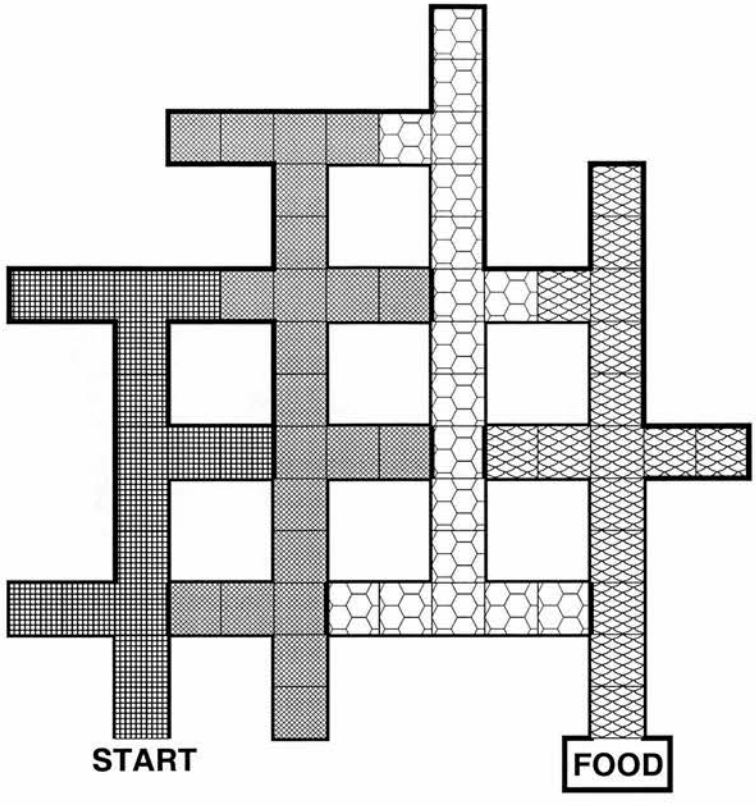


Figure 1.7: A possible structural decomposition into fragments for the complex maze of figure 1.2, which demonstrates that: (1) for navigation between fragments, optimal actions may be specified within the current fragment without knowing exactly where within the goal's fragment the goal is located; (2) fragments may be determined by clustering states on the basis of their values (as shown in figure 1.3).

the maze in figure 1.2. By comparing this figure with the value function in figure 1.3, it can be observed that most of the boundaries between fragments occur at correspondingly large steps in value between states, such as occur when states are on opposite sides of a barrier. However, it is exactly this barrier structure which it is hoped to capture. Therefore, the suggestion explored in chapter 6 of this thesis is that fragments be found by clustering states on the basis of their optimal values, with respect to a subset of goals. The hope is that by learning a better (but still local) representation, generalisation across changes to the navigational task might be coped with using the same reinforcement learning algorithms that have already been considered.

1.7 Plan of The Thesis

To recapitulate, this thesis has two somewhat different concerns. The first is to identify and characterise a possible link between the function of an area of the mammalian brain known as the hippocampus, and a set of general learning algorithms collectively known as

reinforcement learning (RL). Therefore, chapter 2 introduces the hippocampus along with evidence for a role in navigational learning and for a paradoxical role for neurons in this area. Then, chapter 3 reviews RL methods themselves, beginning with the characterisation of navigation as a Markov decision problem (MDP). Ways of solving MDPs are presented: first, a group of methods (dynamic programming methods) which are easy to understand but also implausible from both a computational and neural point of view, then second, a group of methods (model-free methods) which effectively do the same thing but in a more plausible way. Finally, chapter 4 presents a model for how hippocampal neurons could be used in conjunction with RL learning methods to solve both the RMW and DMP tasks.

However, chapter 4 raises other issues, which lead to the second concern of the thesis. A novel scheme is proposed whereby coordinates are learned to support generalisation, and this leads in turn to two predictions. First, because coordinates once learned should be available everywhere in the watermaze, a prediction is made concerning one-trial learning after merely placing rats on a platform located in a novel position. Second, coordinates incorporate the distinct limitation that control is only specified in open, unobstructed environments. Following a review of relevant experimental literature, both predictions are tested in chapter 5, which presents behavioural data for one-trial learning after placement, and in the presence of barriers.

While the first prediction is upheld, the second is not, thus demanding a more general account of generalisation in navigation. Therefore, chapter 6 focuses on achieving generalisation but without recourse to learning complete models of the environment (in the implausible fashion of dynamic programming). Instead, the focus is on learning about the underlying, hierarchical structure of environments, and using this structure to guide learning. Finally, a conclusion brings the various arguments together, and raises some questions for future work.

Chapter 2

The Paradox of Hippocampally Dependent Navigation

The hippocampus is a neural structure that is likely to play both a distinct and fundamental role in navigational learning. The purpose of this chapter is to review the experimental literature relating to this role. First, the hippocampus is introduced anatomically, and then with a very brief list of current and competing descriptions of its function. Then, evidence for a role in navigation is introduced, followed by a discussion of neural mechanisms that might support this role. The conclusion is an apparent paradox that remains to be resolved.

2.1 Introduction to the Hippocampus

2.1.1 The Anatomy of the Hippocampus

The mammalian hippocampal formation is a bilateral structure located in the temporal lobe. It is generally considered to consist of the cornu ammonis regions (CA3 and CA1), the dentate gyrus (DG; also known as fascia dentata), the subicular complex (SC; consisting of subiculum, presubiculum and parasubiculum) and the entorhinal cortex (EC). In this thesis, the following terminology will be adopted: the term *hippocampus* will be used to refer to the CA regions and DG and SC, but not the entorhinal cortex. The term *hippocampal formation* will be used in the conventional sense, to include the entorhinal cortex. In the rat, the combined surface area of the hippocampal formation accounts for approximately 80% of the entire isocortex (Amaral and Witter, 1995).

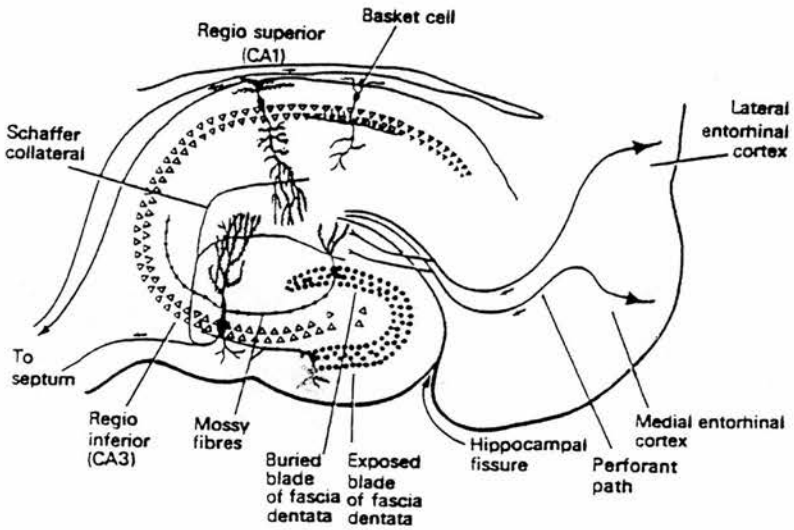
The hippocampus receives its major glutamatergic (excitatory) input from EC, which in turn receives highly processed, multimodal information from several regions of association cortex, including the parahippocampal gyrus (perirhinal and postrhinal cortices), and the parietal, inferotemporal and frontal cortices (Amaral and Witter, 1995). Accordingly, the

range of different kinds of information received by the hippocampus is extremely broad in comparison with much of the rest of the brain. The hippocampus also receives input from subcortical areas (Amaral and Witter, 1995). Cholinergic (modulatory) and GABAergic (inhibitory) inputs arise from the septum and from the nuclei of the diaginal band of Broca, also via the fimbria-fornix. A serotoninergic (modulatory) projection, through the fimbria-fornix, originates in the raphe nucleus. A noradrenergic (modulatory) projection is provided by the locus coeruleus. The hippocampus also receives projections from the amygdala.

A traditional characterisation of the intrinsic circuitry among excitatory pyramidal neurons of the hippocampal formation is as a *feed-forward* circuit of excitatory synaptic transmission, originally referred to as the "tri-synaptic loop" (figure 2.1a): (1) EC pyramidal cells project to DG via the perforant path; (2) DG cells project to the CA3 region via the mossy fibres; (3) CA3 cells project to CA1 via the Schaffer collaterals. However, this description is now widely perceived as an oversimplification (figure 2.1b; Amaral and Witter, 1995). A curious fanning out occurs between EC and DG in that there are many more DG cells than EC cells, and the connections between them are highly divergent. Two pathways lead from the superficial cell layers of EC to CA3 cells, one via DG as above, but another via a direct perforant path projection to CA3. CA3 in fact receives most of its inputs through recurrent connections from other CA3 cells. There is even evidence for a direct perforant path connection from EC to CA1. Nevertheless, no actual feed-back pathways that are glutamatergic (excitatory) in nature have been identified. From CA1 there are two routes back to the deep cell layers of EC: one directly back, the other via the subicular complex which in turn projects back to EC (although it also projects to sub-cortical areas such as the septal complex, the mamillary bodies, ventral striatum and thalamus). The CA1 itself also projects directly to some cortical areas (amygdala, hypothalamus, lateral septum).

Hippocampal anatomy has inspired a number of functional interpretations. The fanning out from EC to DG has been interpreted as a way of orthogonalising input patterns that has useful properties for increasing storage capacity if the aim is to store input patterns (McNaughton and Nadel, 1990), but alternatively as a re-representation of input information for associative learning purposes (O'Reilly and McClelland, 1994). The recurrence in CA3 has been interpreted as an auto-associative memory, for storing memory patterns (McNaughton and Morris, 1987) or memory sequences (Levy, 1996), but alternatively as a way of averaging over navigational actions (Blum and Abbott, 1996). Arguably, hippocampal anatomy is more suggestive than conclusive.

(A)



(B)

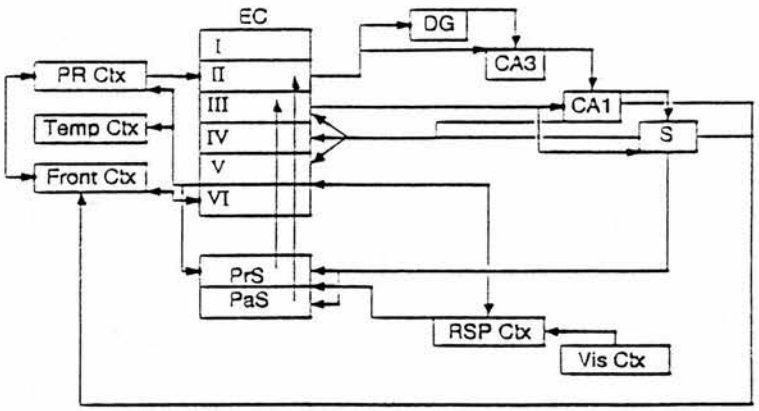


Figure 2.1: (A) A schematic diagram showing a horizontal section through the hippocampus, indicating the subset of intrahippocampal connections comprising the tri-synaptic loop (from O'Keefe and Nadel, 1978). (B) Major intrinsic connections of the hippocampal formation and major extrinsic cortical outputs. This diagram summarises current knowledge of both serial and parallel aspects of the intrinsic hippocampal circuitry such as the direct entorhinal projections to areas CA3 and CA1 not illustrated in (A). Key: PR Ctx, perirhinal cortex; Temp Ctx, temporal cortex; Front Ctx, frontal cortex; Vis Ctx, visual cortex; RSP Ctx, retrosplenial cortex; S, subiculum; PrS, presubiculum; PaS, parasubiculum (from Amaral and Witter, 1995).

2.1.2 Descriptions of Hippocampal Function

The hippocampus has been the subject of intense study, perhaps more so than any other brain area apart from striate cortex. Inevitably, competing descriptions of hippocampal function have been proposed. Each description captures something about the nature of hippocampal function, although each is also, in its narrowest interpretation, wrong. In terms not necessarily similar to those used by the authors themselves, it has been suggested that:

1. The hippocampus is a cognitive map, storing spatial information for the purposes of navigation (**Cognitive Map**; O'Keefe and Nadel, 1978).

2. The hippocampus is a general-purpose, but time-limited memory storage device (**Memory Buffer**; Marr, 1971; McNaughton and Morris, 1987; McClelland *et al*, 1995).

3. The hippocampus is a permanent, "pointer" storage device, to aid the recollection of memories stored extra-hippocampally (**Memory Pointer**; Teyler and DiScenna, 1986; Nadel and Moscovitch, 1997).

4. The hippocampus is a re-representation device, which exports to the neo-cortex representations that transform non-linearly-separable problems into linearly-separable ones, because (it is argued) by itself the neo-cortex cannot solve such problems (**Hidden Layer**; Gluck and Myers, 1996).

5. The hippocampus is a re-representation device, which builds and uses representations of configurations of sensory cues, for the solution of discrimination problems which cannot be solved merely in terms of the presence or absence of individual cues (**Configural Learner**; Sutherland and Rudy, 1989; Rudy and Sutherland, 1995).

6. The hippocampus is a device for discriminating between different behavioural or environmental contexts (**Context Discriminator**; Good and Honey, 1991; Redish, 1997).

7. The hippocampus is a machine for the manipulation of experiential sequences, for the solution of sequential decision problems (**Sequence Manipulator**; Levy, 1996; Bunsey and Eichenbaum, 1996).

One is reminded, perhaps, of the blind men of Indostan describing an elephant. This thesis will focus on the involvement of the hippocampus in navigation, at a computational level of description, which does not preclude compatibility with many of the above characterisations.

2.2 The Hippocampus Is Necessary For Navigational Learning

2.2.1 Preamble

The evidence summarised in this section is the product of a wide range of experimental techniques: lesion studies, pharmacological studies and molecular-genetic studies. The purpose of this preamble is to say a few introductory words about each.

The rationale behind lesion studies follows from the apparently modular nature of the brain (Shallice, 1988). However, brain modules are extremely unlikely to be uniquely allocated to separate *tasks*, but rather to separate *computational processes*, which may each be of use in a great many tasks. Therefore, the efficacy of the lesion approach in revealing the nature of the function of a brain area relies heavily on finding, by design or by luck, tasks which are particularly demanding of the relevant computational process.

Lesion technology has developed over the period during which the following results have been collected. All hippocampal lesions necessitate a certain degree of damage to the overlying cortex. Early hippocampal lesions tended to produce quite diffuse damage. An alternative that was often adopted was to lesion instead the fimbria-fornix sub-cortical input pathway, which is thought to be critical for normal hippocampal function (*eg* Olton and Papas, 1979). However, a critical development was the ibotenic acid lesion, which destroys principal cells without disrupting fibres of passage (Jarrard, 1989).

The means by which neurons in the brain communicate with one another at synapses involves the release of neurotransmitters from presynaptic sites which act upon receptors at postsynaptic sites. Pharmacological studies involve alterations in these processes, often by blocking the action of selected receptors, that can also be localised to particular areas of the brain. In certain cases, the effects of the pharmacological action can even be reversed. Thus these techniques can be considered a more sophisticated set of tools than lesions for investigating the computational function of a brain area (Izquierdo and Medina, 1998). The sophistication does, however, depend on the degree of selectivity of the drug.

Just one focus for pharmacological approaches has been the NMDA receptor, which is likely to be directly involved in mechanisms of *synaptic plasticity* (Martin *et al*, 2000). The NMDA receptor is activated by the same neurotransmitter as is responsible for excitatory *synaptic transmission*, *ie* the normal means by which neurons pass information from one to the other. However, NMDA receptors have the exceptional property that they are only activated when the receipt of this neurotransmitter occurs at the same time as a change in the electrical properties of the postsynaptic site, which generally only occurs when the postsynaptic cell is itself active (Dingledine, 1983; Bliss and Collingridge, 1993; Magee

and Johnston, 1997). Thus the NMDA receptor seems to be activated by the coincidence or correlation of presynaptic and postsynaptic activity. The eventual effect of NMDA receptor activation is thought to be an increase in the effect which the presynaptic cell has on the postsynaptic cell, a phenomenon predicted in essence by Hebb (1949), the computational consequences of which are extremely non-trivial (Dayan and Abbott, 2000). The hippocampus has a particularly high density of NMDA receptors by comparison with the rest of the brain (Monaghan and Cotman, 1985).

There exists an electrophysiologically inducible enhancement effect, referred to as long-term potentiation (LTP), which in many pathways (notably including the perforant path and Schaffer collateral pathways in the hippocampus) requires NMDA receptor activation, and which can be effected in a synapse-specific manner. One important exception in the hippocampus is the mossy fibre pathway from dentate gyrus to CA3, for which LTP is both independent of NMDA receptors, and non-associative, being dependent only on presynaptic activity (Nicoll and Malenka, 1995). A somewhat analogous reduction in efficacy has also been observed, and is referred to as long-term depression (LTD). It is supposed that both LTP and LTD bear a close relationship to biological mechanisms of synaptic plasticity, although there is as yet little direct evidence supporting this supposition (Martin *et al*, 2000).

Powerful new genetic techniques have been developed for interfering with synaptic plasticity in the hippocampus. A recent technological development is the ability, in some cases, to restrict a gene knockout to just one region of the brain (Tsien *et al*, 1996a). Molecular-genetic techniques have been criticised for the abnormal development which a mutant animal must necessarily undergo. However, a second frontier in the field is the development of techniques for switching on and off the function of individual genes in a temporally controlled manner. Therefore, while few results have yet been acquired, this number is likely to grow very rapidly.

2.2.2 Navigation To A Fixed Goal

The RMW task described in the introduction is exquisitely sensitive to hippocampal lesions. Rats with hippocampal lesions cannot acquire the task readily, as normal animals can, and typically perform at chance during a post-training transfer test, in which the platform is removed and measurement made of the amount of time spent within the quadrant in which the platform had been located relative to the other quadrants (Morris *et al*, 1982; Sutherland *et al*, 1983). By contrast, hippocampal lesions do not interfere with learning to navigate to a visible platform in the watermaze (Morris *et al*, 1982). Lesioning the hippocampus of a rat that has acquired an RMW task as normal completely disrupts subse-

quent performance (Weisend *et al*, 1996). Similar results are reported after lesions of the fornix (Eichenbaum *et al*, 1990; Sutherland and Rodriguez, 1990; Packard and McGough, 1992).

Hippocampal lesions effectively restricted to just part of the hippocampus can also interfere with this task. Sutherland and Hoesing (1993) used colchicine, which destroys cells in DG while leaving CA3 and CA1 intact, and found an impairment in RMW performance when rats were lesioned after acquisition. Curiously, this impairment was not evident in animals given colchicine lesions 12 weeks after acquisition training, perhaps suggesting a role for DG in storing only more recently acquired information. By contrast, more complete hippocampal lesions appear to result in impaired RMW performance even 36 weeks after training (Weisend *et al*, 1996).

A navigational apparatus that is in certain respects similar to the watermaze is the circular arena task, also referred to as the Barnes maze (Barnes, 1979). A circular arena is bathed in bright light. In the floor of the arena there are a number of holes, only one of which leads to a dark, escape nest. Rats, prefering the dark, are motivated to find the escape hole and this forms the basis for an analog of the RMW task. Although it is presumably harder to guard against the use of local cues than in the watermaze, the task is likely to be solved through the use of distal cue information. Colchicine lesions impair the acquisition of the task (McNaughton *et al*, 1989).

It turns out, however, that hippocampally lesioned animals *can* learn an RMW task under rather special training conditions. Hippocampally lesioned animals trained with a very large number of trials (*eg* 70 trials) will, on a subsequent trial with the platform removed, search mainly in the same quadrant as the platform has been, to an extent significantly above chance (Morris *et al*, 1990). Alternatively, a "shaping" procedure is effective, as revealed by animals with fimbria-fornix damage, which were first placed on a visual platform, then very close to it, then at the edge of the pool, and only subsequently in a standard, hidden-platform trial (Whishaw and Jarrard, 1996). One possible conclusion from these results is that the hippocampally-lesioned brain is perfectly capable of effecting the mapping from perception to action required to solve the RMW task, but that the difficulty it encounters, and hence the contribution that the hippocampus makes, is in *learning* this mapping. However, given that lesions after learning cause an impairment, a second conclusion is that for at least some period after training, some of the information processing required for navigation is routed through the hippocampus. These conclusions are, of course, perfectly compatible.

Pharmacological treatments have been used to investigate the contribution of hippocampal synaptic plasticity to RMW learning. Morris (1989; see also Morris *et al*, 1986) investi-

gated RMW learning by rats that had been treated with AP5, an NMDA-receptor antagonist, chronically infused into the lateral ventricle. Morris (1989) observed that if AP5 was given during training, the animals failed to acquire the RMW task. If, however, the AP5 was given after a suitable acquisition period, the performance of the animals on the task was unimpaired. Hence, and in contrast to the hippocampal lesion result, learning RMW arguably requires hippocampal synaptic plasticity, whereas performance of a learned RMW task does not.

A potentially awkward result with respect to this was reported by Bannerman *et al* (1995) who pretrained rats on an RMW task in one watermaze, then treated them with AP5 and trained them on a second RMW task in a *different* pool situated in a different room. The rats acquired the second RMW task, despite the fact that hippocampal long-term potentiation, and hence arguably synaptic plasticity, was shown to be blocked. Bannerman *et al* (1995) additionally found that a group pretrained on a "non-spatial" task, in which the platform moved randomly from trial to trial, *were* subsequently impaired at acquiring the RMW task in the second watermaze in the AP5 condition. There are, however, difficulties in interpreting the "non-spatial" group, since AP5 is known to cause perseveration, and so it is possible that the rats in this group revealed perseveration with a random searching strategy rather than a more fundamental learning impairment related to the absence of hippocampal plasticity.

The real difficulty with this result, however, is the issue of the extent of transfer between the two watermazes. It seems to have been generally accepted that tasks in different watermazes would necessarily involve, at some level of representation either in the hippocampus or elsewhere in the brain, separate and distinguishable learning. If this is false, *ie* if learning in one watermaze is greatly facilitated by learning in another, then the result of Bannerman *et al* (1995) resembles the much more straightforward result of Morris (1989) in which acquisition is presumed to have been completed. Two observations are pertinent to this case. First, the data in Bannerman *et al* (1995) does indeed show more rapid acquisition of the second RMW task than the first, by both controls and the spatially pretrained AP5 group. Second, in a study of a DMP task, Whishaw (1991) observed an apparently complete transfer across watermazes, in that after 15 days of DMP training (*ie* 15 novel platform positions) in one watermaze, a single day in a completely novel second pool revealed one-trial learning. These data do not distinguish between the two possibilities of either spatial information being transfered across watermazes (which are in many spatial aspects exactly the same), or simply non-spatial aspects of task performance being so good as to mask the spatial learning that subsequently does or does not happen. Furthermore, the issue of learning in the second watermaze cannot be completely finessed away by the fact of transfer, because even if the rats bring to the second watermaze a learned understand-

ing of everything about the size and shape and navigability of watermazes, they still need to learn how this understanding maps to the unfamiliar sensory information from the new environment. However, this is a greatly reduced learning requirement compared with what the rats must learn about for the first watermaze, and again, their learning difficulties may be masked by positive transfer. At the very least, it can be concluded that the results of Bannerman *et al* (1995) do not support a straightforward interpretation.

NMDA receptors have not been the only focus for pharmacological manipulations in combination with navigational tasks. Cholinergic input both suppresses transmission and enhances plasticity at hippocampal synapses, and has been modeled as playing a key functional role in normal synaptic plasticity (Hasselmo and Schnell, 1994). Studies have shown that a cholinergic blocker interferes with RMW acquisition (Sutherland *et al*, 1982; Whishaw, 1985b). After acquisition, however, the blocker has no effect (Whishaw, 1985b).

Finally, one of the most impressive molecular genetic results so far reported is the following. In a series of experiments, Tsien *et al* (1996a, b) and McHugh *et al* (1996) demonstrated that an alteration in NMDA receptors (deletion of the R1 subunit of the NDMA receptor) specific to just the CA1 subregion of the hippocampus resulted in both alterations which might be interpreted as impairments of the activity of hippocampal neurons as measured electrophysiologically (described in the next section), and also impairments in RMW acquisition. The same manipulation was sufficient to cause a blockade of NMDA-receptor dependent LTP.

2.2.3 Navigation To Multiple Goals

The DMP task described in the introduction is, unsurprisingly, also sensitive to hippocampal lesions. Steele and Morris (1999) first pre-trained rats as normal on the DMP task for 9 days, during which time they acquired one-trial learning, that is, asymptotic navigational performance to the novel platform positions on only the second trial. The subjects were then lesioned, and subsequently training was continued, to another 9 novel platform positions. The one-trial learning performance of the lesioned rats was completely disrupted.

The contribution of hippocampal synaptic plasticity to the DMP task has also been investigated. Over the course of several studies and replications, Steele and Morris (1999) pretrained rats for 9 days (to 9 novel platform locations) as normals before training them in the same pool (to another 9 novel platform locations) under three conditions: under AP5, with hippocampal lesion, or as control. In one experiment, local hippocampal infusion of AP5 was used on intermittant days, allowing each animal to act as its own control, comparing days when AP5 was administered with days when it was not. The delay between the first and second trials of each day was one of 15s, 20min or 2h, varying for each rat

pseudo-randomly from day to day. The results, in terms of trial 2 performance, were as follows: controls were unimpaired at all delays, lesioned animals were impaired at all delays, but the AP5 treated animals were impaired only at the 20min and 2h delays. This suggests a dissociation between two computational components: (1) The *navigational* component of the problem, *ie* the business of working out how to swim directly to the novel platform position, is similar across all delays. Therefore, it can be concluded from the 15s delay that, following extensive pretraining, this component requires the hippocampus but not synaptic plasticity there. (2) The remaining component of the problem, and that which distinguishes between control performance and that of treated animals, is *mnemonic*, in the sense that information about the current platform position must be retained across the delay. Apparently it is this component which, even after extensive pretraining, continues to depend upon hippocampal synaptic plasticity.

Choligernic blockade has also been applied in an earlier version of the DMP task, that used a different but not novel platform on each day (using a limited set of only four possible platform positions). The blockade disrupted the performance of rats that had been pretrained to the point where they showed one-trial learning (Whishaw, 1985b).

2.2.4 The Radial Arm Maze

The radial arm maze was described in the introduction as a potentially rather simpler task than either RMW or DMP from the perspective of navigation. However, the task has been used extensively for the investigation of the effects of various neural manipulations, particularly manipulations to the hippocampus.

The lesion story for the radial arm maze has some similarities to that for the watermaze, but it would be hard to argue the pattern of results is equivalent. It appears that acquisition of the radial arm maze, in terms of both working and reference memory errors, is hippocampal dependent (including evidence from both fimbria-formix lesion and ibotenic acid hippocampal lesion; Sutherland *et al*, 1987a; Jarrard, 1993). However, after extensive pre-training on the task, it appears that hippocampal lesion leads to working memory errors but not to reference memory errors (Olton and Papas, 1979).

A body of work has investigated the contribution of NMDA receptors to various aspects of learning in the radial arm maze. A key issue, again, is the extent and nature of prior experience. The NMDA receptor antagonist MK-801 impairs acquisition of both reference and working memory components of the task (Shapiro and Caramanos, 1990). However, if MK-801 is received only after extensive pretraining, rats continue to perform without impairment with respect to *both* reference memory and working memory components (Shapiro and Caramanos, 1990), a result that appears to be insensitive to the extent of an

imposed delay between arm visits, and to the number of arms to be remembered. This is conspicuously not analogous to the DMP result of Steele and Morris (1999). A further manipulation has been to investigate transfer between radial arm mazes in different environments. Clearly, working memory might be expected to transfer, but not reference memory, and this was found to be the case (Caramanos and Shapiro, 1994) – there is not the evidence to tell whether rats learn the new unbaited arms any quicker than they learned those of the first maze. Using either MK-801 or AP5, Caramanos and Shapiro (1994) additionally found that, after pretraining in one maze, subsequent testing under drug in a second maze revealed reference memory errors (like controls) and working memory errors (unlike controls). It was further reported that reference memory did not improve in treated animals as it eventually did in controls. In summary, the above pattern of impairments appears to be another instance (to go with the lesion results) of a less severe effect of hippocampal disruption in the radial arm maze compared to that in the watermaze.

2.2.5 Summary

The overwhelming weight of evidence suggests that the hippocampus is necessary for navigational learning. Experimental evidence has been cited that used lesions, pharmacological intervention and molecular-genetic techniques. However, certain lines of evidence have been deliberately omitted. One example is the increasing number of functional imaging studies, usually of humans imagining navigational tasks or even performing them in a virtual environment (*eg* Maguire *et al*, 1997). The reason for this omission is simply in accordance with the aims of the discussion – to demonstrate the necessity of the hippocampus for navigational learning. Inevitably, a study which reveals activity in the hippocampus enhanced during a navigational task as compared to a non-navigational one, as indeed many do, fails to show necessity as opposed to mere (even perhaps epiphenomenal) involvement. For the same reason, the electrophysiological study of hippocampal neurons is not treated as evidence for necessity, but is rather discussed in the next section in the context of how the hippocampus might be contributing to navigation. Functional imaging will not be discussed, as the information it reveals is spatially and temporally less precise than the electrophysiological work. This is not to dismiss functional imaging which, like the molecular-genetic technology, is likely to dominate in the future.

A potentially important implication of the pharmacological studies described is that navigational learning *per se* may not always require hippocampal synaptic plasticity. Further, this can be dissociated from the clear requirement for having a hippocampus. This poses difficulties for an explanation of navigational learning in terms of a hippocampal memory for specific navigational paths (*eg* **Sequence Manipulator**). This difficulty in fact mirrors the more compelling paradox of hippocampal neurons, described below.

2.3 The Phenomenon of Hippocampal Place Cells

2.3.1 Preamble

It is possible to observe the firing patterns of individual hippocampal neurons in awake, behaving rats using a technique generally referred to as *single-unit recording*. In its original form, a single electrode tip was inserted into the hippocampus, and the voltage difference across the tip relative to a reference electrode was monitored. By setting a suitable threshold (determined on the fly) it was possible to (1) record individual spikes, *ie* the information-carrying signals emitted from a neuron, (2) listen preferentially to the nearest unit, by raising the threshold to the (hopeful) exclusion of all other units, and (3) use the recorded signal to nestle as close as possible to the current cell. In practice it turned out that, using just one electrode, it was impossible to be sure that only a single neuron was being recorded from. Therefore first the stereotrode (pair of electrodes; McNaughton *et al*, 1983) and then the tetrode (4 electrodes together; Recce and O'Keefe, 1989) were developed, the latter assuring strict isolation. The most recent technological advances include the use of up to 20 tetrodes at the same time (Wilson and McNaughton, 1993). Moreover, the use of data visualisation techniques allows for many more than the minimum of 20 cells to be distinguished, in some cases allowing the simultaneous recording of the activity of hundreds of neurons (Wilson and McNaughton, 1993).

Single-unit recording techniques have been used to characterise a singular feature of the activity of principal neurons in the hippocampus. These are the excitatory neurons of the hippocampus which send glutamatergic projections to other neurons and which are generally thought to be the information-carrying neurons of the hippocampus, as elsewhere. They may be contrasted with the interneurons which send GABA-ergic inhibitory projections to other neurons and are thought to have a less specific role, such as normalising excitatory signals. Single-unit recording revealed that, while a rat occupied an environment, principal neurons acted as *place cells*, that is, they fired only in a restricted part of the environment (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978). Place cells have also been found in SC, although much less is known about their firing properties (Sharp and Green, 1994). The basic properties of place cells in DG, CA3 and CA1 are described in the following section.

2.3.2 First Order Properties of Place Cells in DG, CA3 and CA1

A number of striking properties of place cells have been reported, with many of these properties continually coming under closer scrutiny, and subsequent qualification. An example of why it is important to bear this in mind is afforded by recent observations that a majority

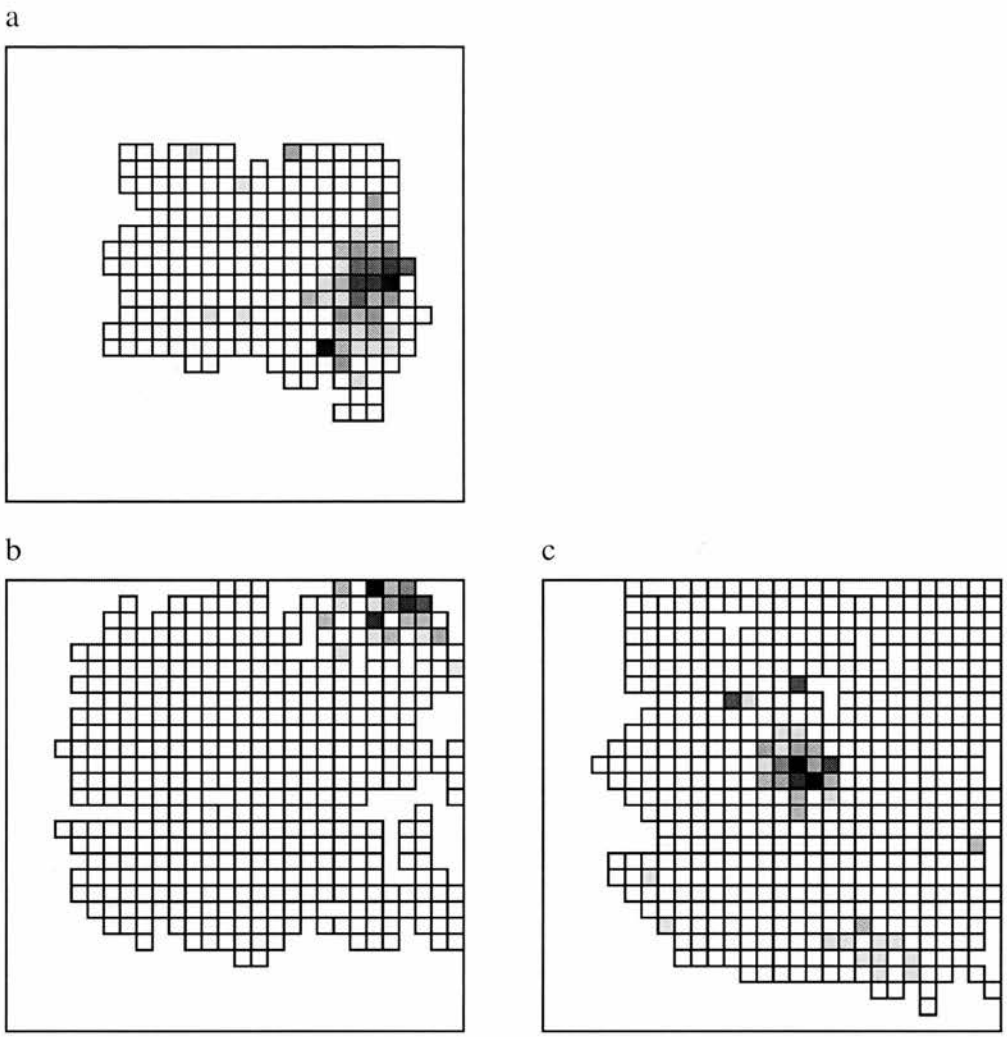


Figure 2.2: Hippocampal place cells are active only in spatially localised regions of an environment: (a) The activity of one hippocampal place cell in an environment of length 1m. The firing rate of the cell, averaged for each pixel over many passes, is indicated in greyscale, with the highest rate indicated by the darkest colour (8.26 Hz). Places that the animal visited in a recording session but in which the cell did not fire are indicated by an unfilled pixel; (b, c) The activity of two cells from a second animal (highest firing rate: for b, 15 Hz; for c, 13.12 Hz). From Dr. Emma Wood, Dept. of Neuroscience, Edinburgh University, with permission.

of recorded hippocampal neurons can, under certain experimental circumstances, show sensitivity to wholly non-spatial factors, and insensitivity to any obvious spatial factor (Wood *et al*, 1999). Therefore, the properties which are currently well documented, and which are found most often, will be listed in this section and referred to hereafter as "first-order properties" (*eg* Redish, 1997). These properties are as follows:

1. Place cells fire in a restricted portion of an environment, referred to as the cell's *place field* (figure 2.2). Place fields are in general localised and unimodal. Cells with non-

unimodal firing fields, sometimes reported as multiple place fields within the same environment, are rare, increasingly so as more reliable techniques for distinguishing individual cells, *eg* tetrodes, have been adopted. It remains uncertain to what extent post-tetrode reports of place cells with multiple fields are due to remaining measurement problems (Wilson and McNaughton, 1993; Wilson, pers. comm.)

2. In many cases, place cells have directionally independent place fields, that is, the firing of the cell does not depend on the direction in which the rat is facing or moving when occupying the place field (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978; Muller *et al*, 1994). Early reports noted that a rat might even picked up and placed in a place field for the corresponding cell to fire (O'Keefe and Nadel, 1978). However, place cells have been found in the subiculum and dorsal presubiculum whose firing is strongly dependent on the direction of traversal of the place field (Sharp and Green, 1994).

3. For any given environment, place fields will cover the entire space, *ie* there is no uncovered area. The firing of an ensemble of place cells with place fields within an environment can be used to predict the location of the rat. Initial reports used fairly rudimentary decoding procedures (Wilson and McNaughton, 1993); more recent reports have achieved increased accuracy (Brown *et al*, 1998).

4. There is no topographic relationship between the site of pyramidal cells in the hippocampus and the location of their place fields in an environment (O'Keefe and Nadel, 1978).

5. As many as 30% of hippocampal pyramidal cells are active in any given environment, implying extensive re-use of place cells in multiple environments (Wilson and McNaughton, 1993). As Biegler (1996) pointed out, a topographic relationship holding across many environments would severely limit the number of environments that could be coded for, or alternatively require an inordinately large number of cells. A distributed code is non-topographic but more efficient.

6. Place fields are acquired within an environment rapidly, although they can tighten up with time. Reports vary from immediate observation of place fields (Hill, 1978) to improvements (tightening up) in the fields over minutes (Wilson and McNaughton, 1993) or even hours (Austin *et al*, 1993). Once acquired, they can remain stable, even after a rat has left the environment to explore another and then returned (Wilson and McNaughton, 1993).

2.3.3 What Do Place Cells Represent?

The questions of what drives the activity of place cells, what factors, environmental or otherwise, control their firing, and what they ultimately represent, have vexed researchers since their discovery.

Early reports stressed the importance of distal, sensory cues in environments in controlling place cell activity. Rotation of a recording apparatus, even with local cues, did not lead to a corresponding rotation of place fields (Miller and Best, 1980). It was, however, recognised that individual distal cues were not themselves necessary, since cues could be removed without affecting place fields, provided sufficient cues remained (O'Keefe and Conway, 1978). However, more recent reports suggest that the distal/local categorisation may not be the correct one. Local cues, such as textures and smells placed along the arms of a cross-shaped maze, can control place fields, as demonstrated by conflicting manipulations to cues such that distal cues are rotated in one direction (*eg* clockwise) and local cues in the other (*eg* anti-clockwise). While some place fields follow the distal cues, still others follow the local (Tanila *et al*, 1997; Shapiro *et al*, 1997). On the other hand, local landmarks in the form of narrow cylinders placed in an arena fail to control place fields unless they are pushed right up against the walls of the arena (Cressant *et al*, 1997). Distal and local as terms are perhaps sufficiently ambiguous to allow a somewhat *ad hoc* allocation in line with these results. However, a more satisfying if less precise explanation would seem to be that place fields follow whatever cues are likely to provide a *stable* frame of reference in the environment. As such, place fields appear to concord with behavioural evidence for a preference for using stable landmarks to guide search for hidden food in an environment (Biegler and Morris, 1993).

A number of studies in carefully cue-controlled environments revealed that place fields can come under the control of individual distal cues. Muller and Kubie (1987) investigated the activity of place cells while rats occupied a small (0.76m diam) circular arena, with surrounding walls occluding extramaze visual cues (0.51m high). The interior of the wall was a uniform grey, except for a white cue card against the wall, as high as the wall, and extending through 100^o. The basic finding was that the place fields of the majority of recorded place cells rotated with the cue card as it was rotated (*ie* moved to a different portion of the wall through an angle of 90^o). Recent work has demonstrated that if two cue cards are used, one white and one black, and rotated independently, the resulting transformation of a place field is a function of the two cue rotations, each weighted by the field's distance from the cue (Fenton and Muller, 1997). In another cue-controlled experiment, O'Keefe and Burgess (1996) used 0.61m high moveable walls to construct four different shapes of rectangular arena in which place cells were recorded (a small square; the same stretched

horizontally; the same stretched vertically; and the same stretched in both directions to make a larger square). Distal cues were present, *eg* a window at one end of the recording room, but because of the height of the apparatus walls such cues were arguably less salient than the apparatus walls; the window might only have served to orient the animals. O'Keefe and Burgess report that most of the place cells fired in all shapes of box, and moreover the location of the place fields maintained a consistent relationship in terms of either fixed or proportional distances to one of the box walls. They also note that in some cases, stretching the walls in a particular direction led to the splitting up of a directionally-independent place field into two directionally-dependent fields in different positions, as if the field in one direction was controlled by the wall opposite to that controlling the field in the other direction. This notion is consonant with theories of directionally independent place field formation through the superimposition of directionally dependent component fields. The ability of cues to control place fields can be contrasted with the inability of a reward location to do likewise. For example, after training rats on a spatial task requiring them to get to a goal, subsequently moving the goal to a different position did not affect the location of place fields (Speakman and O'Keefe, 1991).

One reason to be cautious about interpreting the effects of cue manipulations in terms of direct control of place cells by single cues comes from an experiment in many ways similar to that of O'Keefe and Burgess (1996). Wilson and McNaughton (1993) allowed rats to explore a square environment for 10 min. before being allowed into an adjacent square box which was novel to them, by the removal of a partition between the two boxes. This is clearly equivalent, however, to elongating the original box, and it is at least plausible, in line with the results of O'Keefe and Burgess (1996), that some of the place fields should have been controlled by the wall that "moved" and so should themselves have moved. In fact, this was not observed. Instead, new place fields occured in the novel area, generally supported by place cells that had not been active in the first area. The place fields in the first area remained stable. This rather different pattern of results may be due to two key diffences between the Wilson and McNaughton (1993) experiment and that of O'Keefe and Burgess (1996): (1) Although the height of the walls was similar in both experiments, walls in the former were covered with a variety of distinct visual and tactile cues; this was not the case in the latter. (2) In the former experiment, rats walked themselves across where the partition had been, whereas in the latter experiment the rats were removed and placed back in the apparatus between changes in the arena shape. Clearly these two differences together might have led the rats in the former study to have a better chance of understanding that the box they were in was different than the rats in the latter study, who seem to have been fooled into thinking they were in exactly the same box. The point to be noted, however, is that whichever of the two differences accounts for the different results, each implies a much more complex relationship between sensory cue and place field location than that

previously hypothesised.

A further result that points to a more complex relationship between even just one, salient cue and the locations of place fields has been found by Knierim *et al* (1995). They used an apparatus very similar to that used by Muller and Kubie (1987): a circular recording arena, with a single cue card against the wall, subtending 90^o. However they pretrained animals over several weeks in one of two ways: either by disorienting them each day before transporting them to the arena for a period of foraging, or, in a second group, by giving them the same foraging experience without disorientation. Subsequently, place cells were recorded from while rats foraged, sometimes with the cue card moved (*ie* rotated). Each recording trial was preceded by a disorientation. Note that one might have predicted that the place fields for the group that had been consistently disoriented would have come to rely more heavily on the visual cue. Alternatively, one might have expected the disorientation before recording to have a more effective, disorienting effect on the group that had never before experienced it. Both these expectations turned out to be wrong, however, because the group that was never disoriented during pretraining showed place fields that were controlled by the cue card, but the group that was disoriented during pretraining did not. These results have been interpreted as showing that place cell activity is initially driven by information closely related to the vestibular system, perhaps concerning internal estimates of position, and only later becomes associated with visual and other sensory stimuli, if they have a reliable relationship with the animal's sense of its own position (McNaughton *et al*, 1996). However, a rather important aspect of the Knierim *et al* (1995) result is likely to be the task used, that of foraging for randomly located food pellets. A task making greater demands for allocentric spatial representation might have led to a different result.

A critical complication to the place cell story is the growing amount of evidence suggesting that the very characterisation of these cells as place cells may be misleading. In a recent study, Wood *et al* (1999) trained rats on a delayed non-match to sample task using odours as stimuli. The task was conducted in a test arena in which the rat ran up to a sampling cup filled with sand of a certain odour, and which contained a food pellet at the bottom only if the odour was different to that of the previous trial. Importantly, on each trial the sampling cup occupied a different, unpredictable position within the arena. Wood *et al* found that a majority of the hippocampal principal cells from which they recorded did not display spatial correlates. A large proportion fired instead to portions of the behavioural experience of the animal, *eg* during the beginning of the approach to the cup, or in the middle, or upon arrival at the cup, all regardless of where the cup was positioned. They also found a large number of cells firing to various *combinations* of spatial and non-spatial correlates, such as approach, odour, position and match status, in line with earlier reports (Eichenbaum *et al*, 1987; Wiener *et al*, 1989; Deadwyler *et al*, 1996). Indeed, it has been

proposed that the localised nature of place cell firing represents merely a *combinatorial* dependence on the distances to two or more sensory cues (Burgess, Recce and O'Keefe, 1994; O'Keefe and Burgess, 1996), or upon the bearings of two cues (Wan *et al*, 1993); **Hidden Layer**; **Configural Learner**), and various other combinatorial schemes for localisation are conceivable. Place cells can even be sensitive to the task itself. Markus *et al* (1995) reported that when rats were trained on two tasks, random foraging in an arena, or spatially localised search in the same arena, different subsets of place cells were active in each task, and rapid switching between tasks produced equally rapid switching between place cell representations.

It is important to note that in a different species, macaque monkeys, place cells are not always observed (O'Mara *et al*, 1994). Macaques have a visual system which is both far more akin to that of humans, and far more effective, than that of rats. Hippocampal principal neurons in macaques have been found to respond to views, in a manner in many other respects analogous to that of rat hippocampal place cells (Georges-Francois *et al*, 1999). Such *view cells* are described as firing when the monkey looks towards a certain place, independent of its current position, and often continue to fire after the salient features of the view are obscured, *eg* by drawing curtains in front of a door which has apparently acquired a view response. It is tempting to conclude that hippocampal principal cells in both species encode the elements of an animal's experience, and that it is only the restricted and peculiarly spatial nature of a rat's experience that accounts for the predominance of place cells in that species (*eg* **Memory Buffer**; **Sequence Manipulator**).

For the purposes of this thesis, it matters less exactly how hippocampal principal neurons come to fire the way they fire, than that the characterisation of the way in which they fire should be correct. This is perhaps just as well, because, as has been discussed, a precise account of their dependence on cues, task and behaviour will have to take into account the influence of an animal's beliefs about the nature of the environment. As a broad characterisation of the way in which hippocampal principal cells fire, the following definition is suitable: place cells are a behavioural state space, which individuates (represents individually) situations which an animal finds itself in, in a manner relevant to the current behavioural task. However, of most use will be a specific, quantitative characterisation of hippocampal principal cell activity in navigational tasks, and one such characterisation is considered in section 2.3.6.

2.3.4 Adaptation Of Place Cells

The stability of place cell activity has been noted as a basic property, but under certain conditions adaptation of place cell activity is also seen. This is clearly rather important to

the question of how place cell activity should be modeled. In fact, various aspects of this adaptation contribute to the argument advanced in this thesis.

In open environments, such as a .76m diameter circular arena, place cells have directionally independent firing fields, *ie* the firing of a cell does not depend on the direction in which the rat is facing or moving when occupying the place field (Muller *et al*, 1994). However, in certain circumstances, place fields *are* directionally dependent, *ie* at any given location, different cells fire depending on which direction the rat is travelling in. This has been reported for traversal of the linear arms of a radial arm maze (McNaughton *et al*, 1983), but also during repeated traversal of a linear route within an open arena (Markus *et al*, 1995), suggesting that directionality is more a property of the action history of the animal than of the environment *per se*. It has already been noted that, by manipulating features of an environment, apparently directionally independent place fields can be "separated" into directional components (O'Keefe and Burgess, 1996; section 2.3.3). Recent models have demonstrated how the recurrent connections in the CA3 subfield of the hippocampus could naturally associate directionally dependent cells to create apparently directionally independent responses, without *a priori* information about the positioning of firing fields (Kali and Dayan, 1998; Brunel and Trullier, 1998). In these models, the firing of hippocampal neurons is thought to be *intrinsically* directional, only later appearing to be otherwise, thus suggesting a *rapprochement* with view cell descriptions of neuronal activity in the primate hippocampus (Georges-Francois *et al*, 1999).

A second sort of adaptation that exactly complements directionality is the elongation of place fields along linear routes (Mehta *et al*, 1996). The total activity of the cells among those with fields along the track increased, furthermore the fields shifted backwards along the track, both suggesting the operation of associative mechanisms that might involve recurrent CA3 connections (just as for the establishment of directional indepenedence) (Mehta *et al*, 1996; Blum and Abbott, 1996). A third form of adaptation that has been reported is that place fields do not straddle barriers to movement, and are extinguished if an obstructing barrier is placed within the field, *ie* the corresponding cell ceases to fire in that place (Muller and Kubie, 1987). This effect was found for transparent barriers as well as for opaque ones.

These important properties are not incorporated into the model of place cell firing used in this thesis – because the tasks that will be modeled do not elicit the kind of stereotyped navigational behaviour for which directional fields or field elongation are found, and the tasks do not involve barriers. It does, however, suggest an important conclusion regarding how place cell firing may actually be used. It is difficult to imagine how a **Cognitive Map** of the navigational environment would be served by such behaviour-specific adaptation, or indeed by either directional dependence or the simultaneous representation of successive

locations, under any circumstances. For example, map theories of directionally dependent firing are forced to classify directional firing fields, such as those for the outward and return journeys along a linear arm, into separate *reference frames*, *ie* different maps (McNaughton *et al*, 1996). By contrast, these sorts of adaptation make sense if the purpose of place cell firing is to provide a useful representation for learning actions. *A priori*, the ultimate outcome after following a linear track or route is likely to depend strongly on the direction of travel, and less on the exact position within the track – a useful constraint to incorporate into a representation the role of which lies in learning to choose actions at locations on the basis of eventual outcomes. These assumptions cannot, however, be made for navigation in open environments. Similarly, the actions on either side of a barrier are likely to be different, even though the locations are, in terms of absolute, map-like space, quite similar.

Pioneering studies have examined the effects of disturbing putative mechanisms of synaptic plasticity on the establishment and maintenance of place cell activity, using pharmacological (Kentros *et al*, 1998) or molecular-genetic approaches (McHugh *et al*, 1996). Kentros *et al* report that, in the presence of an NMDA receptor antagonist, apparently normal place fields can form in a novel environment which remain stable for at least 1.5 hours. However, reintroduction of animals into the environment after 24 hours reveals completely new place fields, *ie* the fields that were formed are not stable for as long as 24 hours. Importantly, novel environments experienced as normal give rise to place fields which remain stable in the presence of the NMDA receptor antagonist for more than 24 hours. McHugh *et al* report that a molecular-genetic deletion of the R1 subunit of the NMDA receptor, specific to just the CA1 hippocampal subfield, resulted in impaired place field formation in a novel environment, although place specificity was not altogether abolished. Together with this effect, the animals were also impaired at acquiring an RMW task. The results, then, are somewhat different to those of Kentros *et al*, although differences in experimental approach might well account for the different results, *eg* developmental effects in the McHugh *et al* study. Generally, further results will be needed to clarify the relationship between place cell development and maintenance, and synaptic plasticity.

2.3.5 Head-Direction Cells

Place cells are not the only cells in the hippocampus that fire in an environment-centred fashion. Head direction cells have been found in the rat in the post-subiculum, anterior thalamic nuclei and lateral dorsal thalamic nucleus (Taube, 1995), and in the lateral mamillary nuclei (Leonhard *et al*, 1996), and in primates, in pre-subiculum (Robertson *et al*, 1999). Such cells fire with an almost triangular function of head direction, with elevated firing over a 90 degree range, and with only a single peak. Such peaks are distributed among cells over the full 360 degrees. The representation of head direction is environment-centred: rotating

external cues in turn rotates head direction cells' angular tuning (Taube, 1995); different cells maintain a constant angular relationship to each other throughout any global rotation. This lends some support to models in which the angular relationships between head direction cells are pre-configured (*eg* Zhang, 1996), an argument which has been extended perhaps less convincingly to place cells (McNaughton *et al*, 1996). Head direction cells in the anterior thalamic nuclei have the important property that their firing is suppressed when the rat is not moving, but also if the rat is attempting to move and is being held back (Taube, 1995). One suggestion is therefore that these cells convey a measure of actual movements to the hippocampus, which would be necessary to drive and manipulate the firing of place cells in the absence of sensory input, although the influence on cell activity due merely to being held has not been ruled out as an explanation for the effect.

In summary, robust compass information is likely to be available to navigating animals, and if, as is thought, head direction cells are essentially pre-configured, the compass is likely to be available as soon as an animal begins navigating in an environment.

2.3.6 A Model of Place Cell Activity That Captures What Place Cells Don't Do

Throughout this thesis, the following, extremely simple model of place cell activity will be adopted, that describes this activity as a gaussian function of position. If the rat is at position $\mathbf{s}$, then the activity of place cell i ($1 \leq i \leq N$) is given by:

$$f_i(\mathbf{s}) = \exp\left(-\frac{\|\mathbf{s} - \mathbf{c}_i\|^2}{2\sigma^2}\right) \tag{2.1}$$

where $\mathbf{c}_i$ is the location in space of the centre of cell i's place field, and σ is the breadth of the field, equivalent to the radius of the circular contour where firing is 61% of the maximal firing rate. Gaussian models of place fields have been proposed previously on the basis of experimental evidence (O'Keefe and Burgess, 1996), though not as simple as the radially symmetric form assumed here.

The paradox which it has been the aim of this chapter to make clear is apparent if it is considered how this idealised model of place cells might be used for navigation. Any single cell's spatial tuning suggests a role in navigation, in agreement with studies of damage to the hippocampus, but the activity of the cell, or even of a collection of such cells, simply individuates different locations – it does not directly tell the animal where it is, or where it ought to go. As such, place cells provide a strange sort of **Cognitive Map** (or indeed **Sequence Manipulator**), in which information can only be read about the spot on the map where one currently is. Looking ahead to see which way to go does not appear to be allowed.

An important qualification to this characterisation is to note that a form of sequential replay of place cells does seem to occur, known as the *theta precession effect*. Hippocampal inhibitory interneurons display a synchronised, rhythmic activity known as the theta rhythm. Hippocampal principal cells appear to be affected by this rhythm in that they too display theta activity. Moreover, cells whose place cells have been most recently entered or traversed are active latest with respect to the phase of the theta rhythm. Hence, within a single theta period, there occurs a sort of play-back of a short sequence of place cell activity. However, two points can be noted about this effect. First, it concerns the activity of only a few place cells at a time. Therefore, it cannot be invoked to explain navigational planning in general. Indeed, models of the function of the theta precession effect restrict its use to that of facilitating associations between immediately successive events rather than modeling a more sequence-related role (Burgess *et al*, 1994; Jensen and Lisman, 1996). Second, it is simply not yet known whether theta precession is predictive, or mnemonic, or both (Skaggs *et al*, 1996). The implications, however, are great. If theta precession is predictive, then this implies some knowledge, perhaps embodied in the synaptic weights between place cells, about the spatial relationship between the corresponding place fields. Of course, such predictive firing would be absolutely required for any look-ahead scheme for navigational purposes. On the other hand, a short-term memory trace, which might be useful for forming associations between place fields, would be computationally far easier to implement, requiring no such spatial information. A memory trace is the simpler, and perhaps therefore more likely option, but would not be navigationally useful in the same way.

In the absence of more compelling evidence for flexibly predictive activity amongst place cells, the role of place cells in navigation remains paradoxical. It is suggested by lesion, pharmacological and molecular-genetic studies that place cells are *necessary* for navigation. But how can they possibly be of any use?

Chapter 3

Reinforcement Learning Methods For Solving Hippocampally Dependent Navigation Problems

3.1 Introduction

The purpose of this chapter is to describe a set of computational methods which fulfil a number of criteria: (1) they support an explanation for the limited but necessary role played by hippocampal neurons in navigation; (2) they solve general navigation problems as reinforcement learning problems; (3) they can solve general, sequential prediction problems; (4) they have experimental support, both from behavioural and neural studies.

The chapter is divided into a number of sections. The following section outlines a mathematical problem description, and shows that this description is suitable for navigation and that one component of this description provides a role for hippocampal place cells. The third, fourth and fifth sections describe a set of computational algorithms, beginning with a set of intuitive but implausible algorithms before describing a set of plausible methods which derive in part from these algorithms. The sixth section reviews empirical support for the methods, and for the application of place cells which is proposed. The final, concluding section outlines some of the limitations of the methods.

3.2 Markov Decision Problems

3.2.1 Navigation is a Markov Decision Problem

This section introduces Markov decision problems (MDPs) and relates the elements of MDPs to navigation. A finite, discrete-time MDP consists of four elements:

1. A finite set of *states*, S, often referred to as a state space, which, for navigation, is the set of possible locations which an animal can occupy.

2. A finite set, A, of *actions* available in each state, which, for navigation, is the set of choices of movement from each state, which may include the option of staying put. For convenience, the set of choices of movement is usually considered to be the same from each state; the unavailability of movements due to features of the environment such as barriers is captured instead by the transition function.

3. A *return*, r_t, received at every timestep, which is in general stochastic, its mean given by a reward function, $R(s, a)$, which depends on both the current action (a) and the state (s) it was made in.

4. At every timestep, a *transition* from the current state, governed by a transition function, $\mathcal{P}(s'|s, a)$, which is a specification of the probabilities of moving from one state (s) to another (s') given an action (a).

The key feature of reward and transition functions which must hold for an MDP is the Markov property – that events (*ie* returns or transitions) must not depend on state history, but be governed only by the current state. For any problem, a state space can be constructed that allows rewards and transitions to have the Markov property, if only by including information about the history of the system within the state description. Clearly, however, the effectiveness of the MDP framework is likely to be greatest for problems in which a natural state description allows the Markov property to hold. For navigation, location allows the Markov property to hold.

It is usually only assumed that state space information, as well as the set of possible action choices, is available to the agent or animal solving the MDP. In fact, the punchline for this chapter is that hippocampal place cells may provide a representation of state – because place cells individuate locations without necessarily being informative about the spatial position or significance of locations. However, this representation is distributed, in that a set of place cells is typically active in many locations. *Function approximation* of this sort is not discussed in the context of solving MDPs until section 3.5, and so the further development of this place cell hypothesis is postponed until after that discussion.

It is likewise usually assumed that, for real world applications as well as for animal learning problems, the reward and transition functions are *not* known *a priori*, but have to be estimated, either directly or implicitly in the process of learning a related quantity. However, the first solution methods described in this chapter assume knowledge of both reward and transition functions, and establish intuitions which remain applicable in subsequent, more realistic methods.

The natural focus for navigation is on *absorbing* MDPs of the following form. A single state corresponds to the goal, and paths through the state space terminate there, *ie* the transition function specifies, for all action choices, transitions from the goal state to itself with probability 1. Experience in the MDP consists of a number of *trials*, each of which starts in some state depending on specified starting conditions, and terminates, usually, at the absorbing goal state.

The problem posed by the MDP is to maximise some measure of the total returns. Typically, we shall consider the case of a single, positive return available always for transitions to the goal state, and zero return elsewhere. In this case, a *discount factor*, γ, is used to weight rewards less the more distantly in time they are received. Thus, for each state s, we wish to choose actions in order to maximise:

$$\mathbf{E}[r_0 + \gamma r_1 + \gamma^2 r_2 + ... + \gamma^T r_T | s_0 = s] \tag{3.1}$$

where r_t is the received return at time t, T is the time (varying from trial to trial) at which the absorbing state is first reached, and the expectation $\mathbf{E}[.]$ is over trials beginning in state s. The discounting factor obeys $0 < \gamma < 1$.

Consider a simple navigation task, such as the RMW task discussed in previous chapters, with reward modeled deterministically as $r_t = 1$ for all transitions onto the goal, and $r_t = 0$ otherwise. Because the only time at which there is any reward is at T, equation 3.1 simplifies to: $\langle \gamma^T \rangle$. Therefore, because additionally the constant discounting factor γ is set such that $0 < \gamma < 1$, this expectation is a monotonic measure of the average time it takes to get to the platform from s. Maximising the expectation for each state (*ie* solving the MDP) is equivalent to optimising navigational paths.

3.2.2 What Makes MDPs Difficult

MDPs suffer a set of difficulties that mirror many of those identified for navigation in chapter 1:

Stochasticity

Stochasticity is axiomatic in the MDP framework. It is assumed in the probabilistic definition of both transitions and rewards. It should be noted that stochasticity can exist in both action selection and transitions, *eg* even if actions are selected deterministically, transitions can be stochastic.

Perceptual Ambiguity

This is a solved problem for MDPs, since the assumption of a Markovian state space essentially guarantees sufficient state information. Note also that in the present case an existing state space has been identified – hippocampal place cells in the mammalian brain. As noted in chapter 2, the question of how such neurons come to fire the way they fire is an interesting one, but relatively independent of questions about how they are used. A body of work has looked at the effect on MDPs of perceptual ambiguity problems, but this will not be discussed further (*eg* Singh, Jaakkola and Jordan, 1994; Kaelbling, Littman and Cassandra, 1998).

Temporal Credit Assignment

This is the name given in the MDP literature to the local-from-global search problem identified in the introduction. In MDPs, the quantity we seek to maximise is the total sum of future returns. However, the relationship between the actions that might be specified in an MDP and the expected returns that result is usually unknown. The navigational MDPs considered here make this particularly clear, because a reward is received only at the end of a long sequence of actions. Thus, for any particular trial, some of the actions performed might have been correct (in the sense of maximising equation 3.1), and some incorrect, and there is no information in the end result to distinguish directly between these two sets of actions.

Change

Change, in the context of MDPs, usually implies alterations to the MDP structure in terms of the reward function or transition function. There is increasing interest in methods for dealing with such changes – however, these extensions are in general beyond the scope of the methods discussed in this chapter. One method of responding to changes in the reward function is presented in chapter 4. General methods for dealing with change are discussed further in chapter 6.

Exploration and Exploitation

A further difficulty has been highlighted in the MDP literature concerning the fact that the actions of an animal or agent do not just control how quickly or to what extent the animal or agent obtains reward, they also can determine what the animal or agent can learn about the environment. There is, in effect, a trade-off between *exploration* of the environment,

that is, learning about the world through experience, and *exploitation* of that knowledge in order to maximise returns. Without sufficient exploration, actions will be sub-optimal, but excessive exploration is likely to be undesirable. While this issue is not present in the dynamic programming methods discussed next since all the knowledge that might be explored is in fact assumed, it is of considerable importance in the methods discussed in the subsequent section, and presumably important in all realistic navigational learning situations.

3.3 Dynamic Programming

This section discusses the mathematical technique called *dynamic programming*, or DP. It begins by introducing the typical DP concepts of policy and value function. Then, two types of DP solution are briefly introduced, to illustrate how MDPs can be tackled efficiently. DP techniques typically assume knowledge of the reward and transition functions. This is a considerable degree of prior knowledge, which it cannot be assumed is available to animals. The importance of DP is that more plausible methods can be derived in part from the DP methods, that do not make these assumptions.

3.3.1 Policies

A *policy* is defined as a mapping from states to actions, that is, a specification at each state of what actions to take. This specification may be deterministic (a single action for each state), or probabilistic (a probability distribution over different actions for each state). A stationary policy, π, is one in which the specification $p^{\pi}(s, a)$ of an action $a \in A$ at each state $s \in S$, whether stochastic or deterministic, is the same every time that state is visited. Note that policies implement a form of closed-loop control, which in general implies better control. Note also that policies naturally support a simple form of generalisation: policies do not depend on starting state, even if they are learned during trials with starting states drawn from a restricted set of states (provided they can in fact be learned under such a condition).

An *optimal policy* is one which maximises equation 3.1 *for every state*. For every MDP there exists at least one optimal policy that is also both stationary and deterministic (Ross, 1983). The aim for our methods is to find one such policy.

3.3.2 Value Functions

The key conceptual tool for dealing with the temporal credit assignment problem is the value function. Equation 3.1 defined the measure of total returns without being explicit about the current choice of actions. Value functions are a similar measure of total returns, but are specific to a particular policy. We define the value function $V^{\pi}(s)$ for a stationary policy $p^{\pi}(s,a)$ to be:

$$V^{\pi}(s) = \mathbf{E}\left[\sum_{t=0}^{T}\gamma^{t}\sum_{a\in A}p^{\pi}(s_t,a)R(s_t,a)|s_0=s\right] \tag{3.2}$$

The value function can also be defined without reference to an expectation over specific sequences of states, by instead making use of a *consistency condition* which holds, by virtue of the Markov property, between successive states:

$$V^{\pi}(s) = \sum_{a\in A}p^{\pi}(s,a)\left(R(s,a)+\gamma\sum_{s'\in S}\mathcal{P}(s'|s,a)V^{\pi}(s')\right) \tag{3.3}$$

The *optimal value function*, $V^{\star}(s)$, can be defined by a consistency condition known as the Bellman equation:

$$V^{\star}(s) = \max_{a}\left(R(s,a)+\gamma\sum_{s'}\mathcal{P}(s'|s,a)V^{\star}(s')\right) \tag{3.4}$$

This in turn allows a deterministic optimal policy, $\pi^{\star}(s)$, to be defined as:

$$\pi^{\star}(s) = \arg\max_{a}\left(R(s,a)+\gamma\sum_{s'}\mathcal{P}(s'|s,a)V^{\star}(s')\right) \tag{3.5}$$

There are two ways in which an optimal policy may be found without already knowing the optimal value function, and these are described in the next two sections.

3.3.3 Value Iteration

A value iteration algorithm is as follows:

1. Initialise $V(s)$ arbitrarily, for all s.

2. For all s and a, calculate $Q(s,a) := R(s,a) + \gamma \sum_{s' \in S} \mathcal{P}(s'|s,a)V(s')$.

3. For all s, set $V(s) := \max_a \{Q(s,a)\}$.

4. Go to (2), unless values have not changed (or have changed by less than some predefined amount).

This algorithm is guaranteed to result, after sufficient iterations, in an arbitrarily close approximation to the optimal value function (Bellman, 1957; Puterman, 1994; Bertsekas, 1995). The order in which states are updated may be arbitrary, *eg* with at intermediate stages some states having been updated more than others, provided states are updated often enough. From the resulting optimal values, an optimal policy may be extracted, which is also stationary and deterministic, using equation 3.5. Value iteration is the basis for an algorithm called Q-learning, discussed later in this chapter.

3.3.4 Policy Iteration

A second approach to determining an optimal policy is to work in the space of policies. The policy iteration algorithm is as follows:

1. Start with an arbitrary, stationary, deterministic policy $\pi(s)$.

2. Compute $V^\pi(s)$ by finding a solution to the linear equations $V^\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s,\pi(s))V^\pi(s')$.

3. Compute an *improved* policy $\pi'(s) = \arg\max_a \{R(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)V^\pi(s')\}$.

4. If $\pi'(s) = \pi(s), \forall s$ then end, otherwise set $\pi = \pi'$ and go to (2).

This algorithm is guaranteed to result, after sufficient iterations, in an optimal value function and optimal policy (Ross, 1983), and in practice often takes fewer iterations to do so than value iteration (Kaelbling *et al*, 1996).

Consider again a simple navigation problem, in which, as has been established, the value function reports a function related to the distance of a state from the goal. Policy iteration works because an action which leads to a large increase in value is guaranteed to take the animal closer to the platform. More generally, policy iteration relies on the Markovian nature of the value function – it has all of the future and none of the past – and so (for a correct value function) there is never the worry that an immediate gain might be offset by some loss in the future. In this way, values provide immediate information about actions of the sort that the actual instantaneous returns by themselves fail to provide.

3.4 Model-Free Methods For Solving MDPs

Having considered the problem of finding an optimal policy given a complete MDP – including reward and transition functions – it should now be acknowledged that an animal generally will not have *a priori* knowledge of either. Reinforcement learning (RL) is the study of the general problem of learning to optimise control in MDPs without assuming *a priori* knowledge of this sort.

Within RL, there is a debate between two positions. On one side are those advocating the learning and use of *models* of the MDP, by which one usually means learning and representing explicitly the reward and transition functions, so that an optimal policy may be computed using DP or similar methods. However, learning the whole transition function is potentially wasteful, since it is likely to involve as many as $|S|^2|A|$ parameters, as opposed to the $|S||A|$ parameters needed to represent the resulting policy (Watkins, 1989). Even more importantly, the task of learning an entire transition function imposes an enormous exploratory burden – not only would the sampling of all state transitions take an extremely long time, but the required behaviour may be incompatible with ongoing navigational demands, *eg* the need to get quickly to the current goal in either the RMW or DMP task. Given a current task (*ie* a specific reward structure), an agent can explicitly control its own exploration and learn an approximate model sufficient for the current task, a strategy which brings model learning closer in efficiency to model-free methods, albeit at significant computational cost (Kearns and Singh, 1998). It remains remarkable, however, that far simpler methods of learning can be just as efficient.

The alternative to model-based methods generally involves learning parametrised representations of the value function and policy *directly*. Information about the reward and transition functions is substituted for by samples of returns and transitions from the environment. This thesis chooses to focus on this group of methods not only because of their efficient use of resources, but also because what little experimental evidence there is for the use of RL methods in animal brains clearly favours model-free approaches. Both aspects of the choice will be discussed in this chapter.

However, even without direct experimental support, model-free methods might be favoured for the following additional properties:

1. They involve only *local* computations, that is, the algorithms use information that is available to an animal at the current state, or as a locally updateable function of states that it has visited. This is in contrast to the dynamic programming methods of the previous section which required the solution of the entire MDP before a single action could be specified.

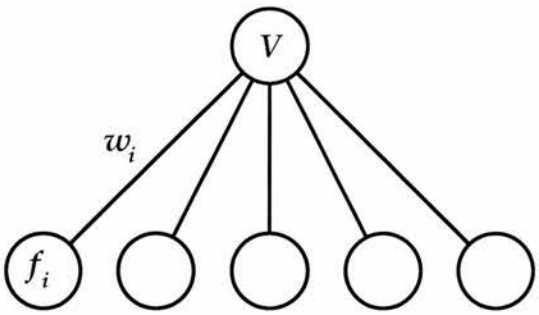


Figure 3.1: A simple connectionist architecture for the approximation of a value function. The input units provide a representation of state, of the form $f_i(s)$. The value is represented as a weighted sum of these inputs, $U(s) = \sum_i w_i f_i(s)$.

2. They are *reactive*, that is, they are always able to specify an action immediately, and do not require a lengthy process of calculating the appropriate action. Dynamic programming techniques require a lengthy process to solve MDPs even once the transition function is known, during which no actions can be specified.

3. They are *incremental*. If DP techniques were to be applied to animals, a very long period of ignorance would have to be assumed, during which the transition function was being learned through trial and error estimates of transition probabilities, and during which actions could be specified only randomly. Much evidence suggests that animals in fact incrementally improve performance in many tasks – so that even when learning is not complete, fairly good actions can be specified.

3.4.1 A Connectionist Framework

Model-free methods can be readily expressed in terms of connectionist learning rules and architectures. A formal neuron is defined as a unit with a real, scalar output, computed at every timestep. Figure 3.1 shows a simple architecture using such neurons. The first layer of neurons consists of neurons whose outputs are predefined functions of state, shown as $f_j(s)$, $j \in 1, ..., N$. The second layer neuron has an output given by the sum of the first layer outputs weighted by the learned parameters w_j, that is, $\sum_{j=1}^{N} w_j f_j(s)$.

The architecture as defined supports, for example, the approximation of an estimated value function, as a function of state. The very simplest scheme is a look-up table representation, in which each state s is uniquely associated with a neuron i in the first layer; when state s is occupied, the associated neuron has an output of $f_i(s) = 1$, and all the rest have an output of $f_{j \neq i}(s) = 0$. The second layer neuron naturally computes the current estimate of the value function $U(s) = \sum_j w_j f_j(s)$. Clearly in this case, individual weights are equivalent to the estimated values of states. This case is assumed throughout the following discussion. However, more complex representations are possible.

An architecture with several units in the second layer can support the approximation of a policy, for example by having each second-layer unit associated with the system's preference for a particular cardinal action, such as north, north-east, etc. This case is described in section 3.4.6, and examined further in the model of chapter 4.

3.4.2 Temporal Difference Learning

The following sections describe methods for learning values, given a *fixed*, possibly stochastic policy. One use for these techniques is *evaluation* of a current policy, as part of a scheme for improving the policy, such as policy iteration. An equivalent situation is that of a stochastic Markov process which cannot be altered, but in which one wishes to make *predictions*.

Consider a fixed but stochastic navigational policy π, in which trials of duration T end at a rewarded goal, where T varies from trial to trial, and the starting state s varies from trial to trial.

One way of learning values is to estimate them directly, by repeatedly sampling the full sequence of returns that happens after occupying each state. Implementation of this idea in a connectionist framework leads to a form of *supervised learning*, in which actual outcomes act as teaching signals for value learning (Sutton, 1988).

An outcome for each trial can be defined as $z = \sum_{t=0}^{T} \gamma^t r_t$ (where $s_0 = s$). The following weight update rule, applied after each trial, minimises the *difference* between the outcome and the estimated value:

$$\Delta w_j \;=\; \eta\,(z - U(s))\,f_j(s) \tag{3.6}$$

where $U(s) = \sum_j w_j f_j(s)$, s is the starting state for the trial, and η is a learning rate. Provided η is not too big, $U(s) \rightarrow V^{\pi}(s)$ for all s (Widrow and Hoff, 1960).

Two criticisms have been made of this approach (Watkins, 1989). First, the sampled outcome z is available only after time T. This is, in effect, a memory requirement, demanding that reward information is cached during this period. Second, the variance in the values of z is likely to be extremely high – because of the combinatorial explosion in possible sequences of action choices, some of which result in far longer paths to the goal than others. This is the very problem raised in the introduction, and which the dynamic programming methods of the previous section address.

Temporal difference (TD) learning learns value estimates not by minimising the difference between predictions and outcomes directly, but by minimising the difference between suc-

cessive predictions. As such, the idea has been around for some time (*eg* Samuel, 1959). The form of TD rule considered here is the general form presented by Sutton (1988) and by Barto, Sutton and Watkins (1990). It addresses both criticisms of supervised learning.

The TD learning rule applied to the value approximation architecture prescribes a way in which the weights are changed to make the estimated value function correct. A quantity called the *prediction error* is defined at each time step as:

$$\delta_t \quad = \quad r_t + \gamma U(s_{t+1}) - U(s_t) \tag{3.7}$$

where r_t is the return received at time step t, and $U(s)$ is the current estimate of the value of state s, as given by the output of the second layer neuron in the network.

The prediction error is intimately related to the recurrent definition of the value function in dynamic programming, equation 3.3, which specifies the value $V^\pi(s)$ of a state s, given a policy π. The implication of this equation is that:

$$\left(\sum_a p^\pi(s,a)R(s,a)\right) + \left(\gamma \sum_a p^\pi(s,a) \sum_{s'} \mathcal{P}(s'|s,a)V^\pi(s')\right) - V^\pi(s) \quad = \quad 0$$

This consistency condition holds for the true value function $V^\pi(s)$. However, for an estimated value function $U^\pi(s)$, the consistency condition may not hold. In this case, the following quantity resembles an error:

$$\left(\sum_a p^\pi(s,a)R(s,a)\right) + \left(\gamma \sum_a p^\pi(s,a) \sum_{s'} \mathcal{P}(s'|s,a)U^\pi(s')\right) - U^\pi(s) \tag{3.8}$$

Without knowledge of the reward or transition functions, this quantity cannot be calculated. However, every time an animal makes a move from state s (at time t), it is able to obtain both a quantity r_t that varies but whose mean is $\sum_a p^\pi(s,a)R(s,a)$, and a quantity $U(s_{t+1})$ that also varies but whose mean is $\sum_a p^\pi(s,a) \sum_{s'} \mathcal{P}(s'|s,a)U^\pi(s')$. Substitution of these quantities into equation 3.8 gives the prediction error, δ_t, which can be applied within a standard connectionist learning rule.

By analogy with equation 3.6, the prediction error is used to specify a change to each weight at each timestep:

$$\Delta w_j \quad = \quad \eta \delta_t f_j(s_t) \tag{3.9}$$

where η is the learning rate. Various convergence results apply (section 3.4.4).

The TD rule allows updates at every step which involve only temporally local information. This removes the memory requirement of the supervised learning rule described in the previous section. The second criticism of supervised learning, that outcomes are likely to be extremely variable, is also addressed by TD learning (section 3.4.5).

3.4.3 TD(λ)

The TD rule of equation 3.9 considers predictions across a single timestep. In general, however, it might be better to make use of predictions across several timesteps. One way of achieving this is to use the TD(λ) algorithm, which is a generalisation of the TD rule of equation 3.9 (Sutton, 1988).

TD(λ) specifies weight changes at each timestep according to the following update rule:

$$\Delta w_j = \eta \delta_t \sum_{k=1}^{t} (\gamma\lambda)^{t-k} f_j(s_k) \tag{3.10}$$

where δ_t is the same prediction error as before (defined by the current time step only), and $0 \leq \lambda \leq 1$. The sum term would seem to render the rule no longer locally implementable, but in fact the algorithm can be implemented locally in a very straightforward manner, by making use of an activity trace. The rule can be re-expressed as:

$$\Delta w_j = \eta \delta_t g_j(t) \tag{3.11}$$

where the *eligibility* $g_j(t)$ is set initially (at time $t = 0$) to zero, and updated at each timestep as follows:

$$g_j(t+1) = f_j(s_{t+1}) + \gamma\lambda g_j(t) \tag{3.12}$$

In this way, TD(λ) considers simultaneously many consistency conditions, across many timesteps, but further separated estimates are weighted with exponentially less importance than closer ones.

An alternative way of looking at the TD(λ) learning rule has been presented by Watkins (1989), in terms of estimates of returns. Note that the following discussion assumes a

look-up table representation of state, specifying changes to the value of a state directly, rather than changes to weights in a connectionist approximation network. Watkins defines a quantity called the "corrected n-step truncated return" $r_t^{(n)}$:

$$r_t^{(n)} = r_t + \gamma r_{t+1} + ... + \gamma^{n-1} r_{t+n-1} + \gamma^n U_{t+n}(s_{t+n}) \tag{3.13}$$

where the correction is provided by the current estimated value function $U_{t+n}(s)$. The estimate of returns relevant to TD(λ) is given by a combination of these corrected truncated returns, across all values of n, with larger values of n weighted exponentially less according to the factor λ:

$$r_t^{\lambda} = (1-\lambda)[r_t^{(1)} + \lambda r_t^{(2)} + \lambda^2 r_t^{(3)} + ...] \tag{3.14}$$

Watkins argues that such an estimate of returns is likely to be an improved estimate of value over the current estimates $U(s_t)$, and so this estimate of returns provides a suitable target for an error term. Watkins considers changes to the estimated value proportional to this term, showing that:

$$\begin{aligned} \Delta U(s_t) &= \eta[r_t^{\lambda} - U(s_t)] \\ &= \eta[\delta_t + \gamma\lambda\delta_{t+1} + ...] + \\ &\quad \eta^2[\gamma\lambda g(s_t, t)\delta_t + (\gamma\lambda)^2 g(s_{t+1}, t+1)\delta_{t+1} + ...] \end{aligned} \tag{3.15}$$

where $g(s,t)$ is the eligibility of state s at time t, defined purely in terms of states, *ie* $g(s,t) = \sum_{k=1}^{t} (\gamma\lambda)^{t-k} f(s,k)$ where the state function $f(s,t)$ is equal to 1 if $s = s_t$ and 0 otherwise. This equation demonstrates that an approximation to the desired error term is readily computable in terms of *local* prediction errors (the first term on the R.H.S. of equation 3.15), the undesirable extra term (the second term on the R.H.S. of equation 3.15) being negligible for a small enough choice of learning rate η. TD(λ) specifies weight changes proportional to the first term, and so is likely to improve the estimated value function.

Two interesting special cases of the TD(λ) rule can be considered. If $\lambda = 0$, *ie* TD(0), equation 3.10 simplifies to equation 3.9, and predictions are considered across a single step only.

If $\lambda = 1$, *ie* TD(1), all ranges of prediction are considered equally. However, a more insightful explanation of TD(1) comes from considering what happens after sampling a

full sequence of returns. All the terms cancel except for the first and last value estimates, and all the returns. Hence, if the value of the absorbing state is fixed at zero, TD(1) is actually equivalent to the supervised learning algorithm. That is:

$$\begin{aligned}\sum_{t=0}^{T} \Delta w_j &= \sum_{t=0}^{T} \left(\eta \left[r_t + \gamma U(s_{t+1}) - U(s_t) \right] \sum_{k=1}^{t} (\gamma)^{t-k} f_j(s_k) \right) \\ &= \eta \left(\sum_{t=0}^{T} \gamma^t r_t - U(s) \right) f_j(s)\end{aligned}$$

where $s_0 = s$.

3.4.4 Convergence Results

TD learning methods, and TD(λ) in particular, are much easier to prescribe than to understand, and the acquisition of useful convergence results has been a gradual process. Sutton (1988) proved that TD(0) converged in the mean to the correct predictions, Dayan proved that TD(λ) also converges in the mean (Dayan, 1992) and, more reassuringly, Dayan and Sejnowski proved that TD(λ) converges with probability 1 (Dayan and Sejnowski, 1994). However, the most comprehensive proof to date is due to Tsitsiklis and Van Roy (1997) who have proved that TD(λ) converges with probability 1, not just with look-up table representations but with function approximation that reduces the number of learning parameters, to within a bounded approximation error. This result is discussed further in the context of function approximation (section 3.5).

3.4.5 Bootstrapping in TD Learning

One might expect that because TD(1) is based solely on actual returns, it would be the rule of choice for most learning situations. However, this expectation turns out to be wrong. In general, lower values of λ are desirable, converging more quickly to better predictions. Although establishing this as an analytical result has proved difficult, several demonstrations exist in the literature. An intuitive understanding for why this might be so can be obtained by considering the variance of predictions in a sequence, and a learning phenomenon generally referred to as "bootstrapping".

The term "bootstrapping" derives from one of the tales of Baron von Munchausen, in which, finding himself trapped at the bottom of a deep well with no prospect of climbing out, the Baron was forced to pull himself out of the well by his own bootstraps (or bootlaces). Bootstrapping in learning begins with the recognition that learning processes often

occur over a period of time, and that, while many learning algorithms call for periods during which all new information is treated in exactly the same way (as if it had been available in parallel), more efficient learning processes should continually incorporate information that is already known with new information.

TD learning is an example of bootstrapping, in the sense that predictions are evaluated in terms of other predictions. This can be understood in terms of trading off bias against variance (Watkins, 1989). With stochasticity governing the transitions between states, direct estimates of the distal effects of immediate actions are likely to be extremely variable. In contrast, the results of single transitions are likely to be rather less variable, and so TD methods (with $\lambda < 1$) might be achieving efficiency by concentrating on achieving local consistency. Working against this effect is bias, *eg* in the extreme case of $\lambda = 0$ and with a look-up table representation, at least one sample of the entire path is required to send information about rewards back across each transition, which would clearly be wasteful if actions and transitions were deterministic. In fact, for a range of stochastic prediction problems, intermediate values for λ have proved best (Sutton, 1998; Singh and Dayan, 1998; Sutton, 1999). It is even likely that a schedule for altering λ during the course of learning is better than any single setting, although prescriptions for exactly how to do this have been conflicting (*eg* Watkins, 1989; Tsitsiklis and Van Roy, 1997).

3.4.6 Learning Actions With TD Learning: The Actor-Critic Architecture

Policy iteration suggests a means by which value functions for a given policy can be used to specify an improved policy. An analogous connectionist computation can be prescribed, called the actor-critic.

A stochastic policy can be represented using a similar architecture to that for value functions, in which a first layer state representation maps to a second layer action representation. The ith second layer neuron represents the relative preference for action a at state s, $\rho_a(s)$, as a weighted function of the outputs of first layer neurons:

$$\rho_a(s) = \sum_j w_{aj} f_j(s) \tag{3.16}$$

where w_{aj} is the weight from input unit j to action unit a. These action preferences support stochastic generation of action choices. A common scheme is to use the soft-max distribution (Bridle, 1989; Nowlan, 1991), which specifies the probability of choosing action a as follows:

$$p(s,a) = \frac{e^{\beta \rho_a(s)}}{\sum_{a'} e^{\beta \rho_{a'}(s)}} \tag{3.17}$$

where β is an inverse temperature parameter which effectively governs the degree of stochasticity in the choice. For example, if β is high, the most preferred action will almost always win, but as β is decreased so the probability is increased of choosing less preferred actions. It may be asked why stochasticity in action choice is ever useful, given that an optimal policy is guaranteed to come from within the subset of deterministic policies. Clearly, however, at early stages of learning, sufficient exploration is required to learn about the effects of different actions. The above stochastic specification of action allows for a smooth change between a situation in which actions are fairly equally preferred (*ie* none are preferred) and so stochasticity is useful, and a later situation in which particular actions are strongly preferred, and action choice is effectively deterministic. This is one way of controlling the "exploration/exploitation" trade-off.

Policy iteration requires the current policy to be evaluated, then a new policy is specified using a maximisation step (section 3.3.4). Policy learning in the actor-critic accomplishes much the same in a stochastic, connectionist context using a correlational learning rule weighted by the prediction error, δ_t:

$$\Delta w_{aj} = \eta^A \delta_t f_j(s_t) g_a(t) \tag{3.18}$$

where η^A is a learning rate for the actor, and $g_a(t)$ is the eligibility of action unit a at time t. In the simplest case, $g_a(t) = 1$ if action a was chosen at time t and $g_a(t) = 0$ otherwise. However, generalisation between action units can be supported by using a more distributed representation.

The actor-critic works because if values are correct, then the mean value of δ_t will be 0. On those occasions when $\delta_t > 0$, the chosen action resulted in a *better* value than the mean value expected under the current policy (*ie* $r_t + \gamma U(s_{t+1}) > U(s_t)$), and so the choice of that action at the current location should be reinforced. The converse should happen if $\delta_t < 0$.

Clearly, the learning in the actor (equation 3.18) and learning in the critic (*eg* equation 3.9) can be made to occur concurrently. However, policy iteration in DP requires values to be determined for a current policy before changes to the policy are made. Similarly, it has been suggested that actor learning should proceed much more slowly than critic learning, *ie* $\eta^A << \eta$ (Witten, 1977). However, in practice this condition is often unnecessary.

Few theoretical guarantees exist for the success of actor-critic learning, although Sutton and his colleagues have recently shown that a form of policy improvement with a separate policy function approximation scheme is convergent to a locally optimal policy (Sutton *et al*, 1999). A related algorithm for which theoretical guarantees are available is described in the next section.

3.4.7 Q-Learning

Just as the actor-critic with TD learning is a model-free version of policy iteration, there is also a model-free version of value iteration, Watkins' Q-learning (Watkins, 1989). Q-learning handles both reward prediction and policy maximisation aspects within the same representation. We can define $Q^\star(s,a)$ as the expected, discounted sum of future rewards if action a is taken for one step, and an optimal policy followed thereafter. Given that the optimal value of a state is $V^\star(s) = \max_a Q^\star(s,a)$, it follows that:

$$Q^\star(s,a) \quad = \quad R(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \max_{a'} \left\{ Q^\star(s',a') \right\} \tag{3.19}$$

Optimal values clearly define a deterministic optimal policy $\pi^\star(s)$:

$$\pi^\star(s) \quad = \quad \arg\max_a \left\{ Q^\star(s,a) \right\} \tag{3.20}$$

Q-learning considers current estimates of Q-values, $Q(s,a)$. Just as in TD learning, samples are substituted for unavailable quantities. For a given state, s, and action, a, the resulting return and next-state can be sampled, yielding s' and r. The following update rule can then be applied:

$$\Delta Q(s,a) \quad = \quad \eta \left(r + \gamma \max_{a'} \left\{ Q(s',a') \right\} - Q(s,a) \right) \tag{3.21}$$

This is *one-step Q-learning* (Watkins, 1989). If each action is executed in each state an infinite number of times on an infinite run and η is decayed appropriately, the algorithm converges to optimal Q-values with probability 1 (Watkins 1989; Watkins and Dayan, 1992; Jaakkola *et al*, 1994). A form of multi-step Q-learning, analogous to TD(λ), has also been proposed (Peng and Williams, 1994).

An important feature of Q-learning is that it will converge appropriately independently of the order in which samples of transitions and rewards are collected. In particular, it is not always necessary for the best current policy to be followed. This flexibility it similar to that of value iteration, for which states can be updated in arbitrary order.

There is no general, computational reason for always favouring either Q-learning or the actor-critic. However, the neurobiological evidence such as it is appears to favour the actor-critic, as discussed later on.

3.4.8 A Navigation Example

So far, a variety of simple methods has been presented for solving MDPs. The previous section laid the groundwork for this section by introducing dynamic programming, and in particular the concepts of policy and value function. However, dynamic programming assumes *a priori* knowledge of quantities such as the reward and transition functions of the MDP. The present section has described model-free methods for solving MDPs without such *a priori* knowledge. In particular, the TD algorithm was introduced as a way of learning values, along with a scheme for using these values in a manner akin to policy iteration and known as the "actor-critic". Note, so far all the methods have been presented as using the simple state representation scheme of one weight per state, often referred to as a "look-up table".

As a demonstration, consider the actor-critic and TD learning applied to an MDP resembling the RMW task described in chapters 1 and 2. The state space of 316 states is circular and discretised, with a diameter 20 states long, and an absorbing goal state at (6,6). The transition matrix is taken for simplicity to be deterministic given a choice of action – from north, south, east or west – which only fails if a wall gets in the way, in which case the current state remains the same. Note that the layout of this "environment" roughly corresponds to the watermaze, in which the platform (11cm diameter) is roughly 1/20 the length of the pool (200cm diameter). Transitions onto the goal state return a reward of 1, and all other transitions are unrewarded.

The actor chooses actions according to equations 3.16 and 3.17. The exploration parameter β was set to 1. The critic alters its value function (*ie* the value weights, w_j) using the prediction error, δ, given by equation 3.7, within the rule given by equation 3.9, and the actor likewise changes its action weights w_{ij} using equation 3.18. The learning rates were set at reasonably well optimised values ($\eta = 1.0$; $\eta^A = 4.0$). The discount factor, γ, was set to 0.9.

Figure 3.2 shows an acquisition curve generated in the following manner. For 10 independent runs, at the beginning of which all weights were set to zero, the system was given 1000 learning trials, each starting from one of four positions at the north, west, south or east edge of the space (thus resembling standard watermaze protocol). Every 5 trials, the numbers of steps taken by the system were recorded for a set of 400 non-learning trials (100 from each starting position). The figure shows means over the 10 runs (the standard error in the means are not shown but the largest was 7.58, and beyond trial 380, all standard errors were less than 1). The system acquires optimal performance (indicated in the figure by the solid line). A typical value function at trial 1000, as shown in figure 3.3, provides a reasonable measure of distance from the goal location, although value information is informative

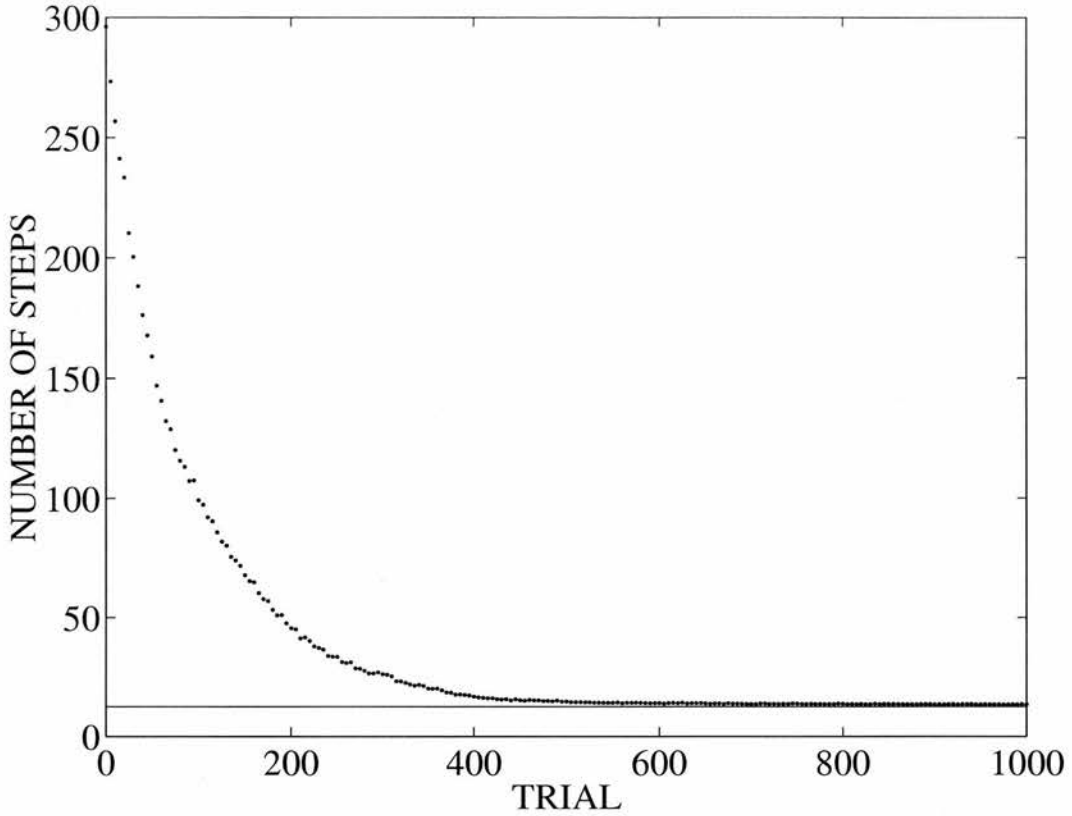


Figure 3.2: TD learning in an actor-critic can learn to solve a simple navigational MDP of 316 states, but the actual acquisition rate can be slow if a look-up table representation is used, as here.

mainly along the direct paths from the four starting locations. Note that the algorithm does not specify a value for the goal state itself, and so this has been set to 1.

This example makes two points. First, TD-based learning methods can clearly solve navigational optimisation problems of this sort. Second, however, the number of trials required to solve such a problem is clearly too large (around 400 trials) to account for the acquisition rates of animals. Certainly using a smaller number of states would in general speed up learning – but here there is the constraint of the "size" of the platform (relative to the size of the environment), and using states much larger than this would presumably mean that the goal state would not be guaranteed to lead to reward. The solution of this problem is to use distributed representations of state – to achieve state accuracy but with fewer parameters. In fact, just such a representation is suggested by the activity of hippocampal place cells. So the next section examines the use of function approximation in reinforcement learning.

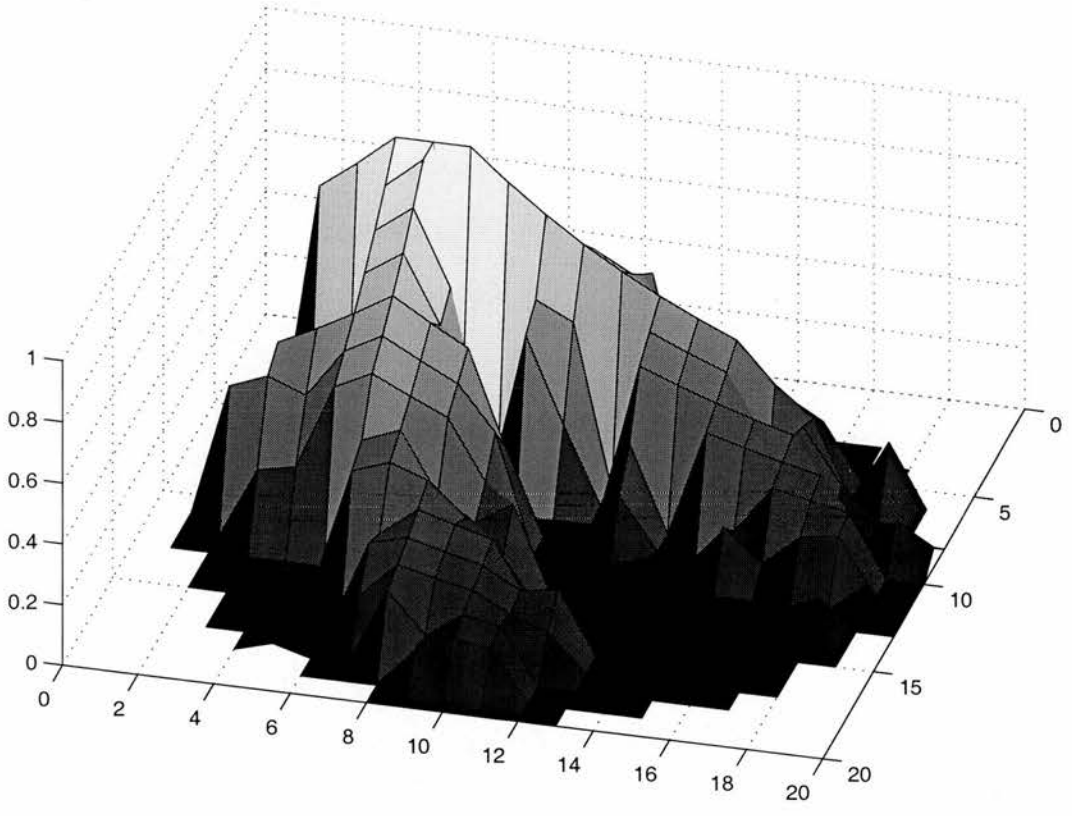


Figure 3.3: The value function for the simple 20×20 navigational MDP after 1000 trials. The peak is clearly visible over the goal state at (6,6).

3.5 Function Approximation

TD learning has been introduced as a connectionist learning rule within a simple network architecture for the approximation of a value function. The discussion has focused on the case where weights were uniquely related to individual states (the "look-up table" case). This begs the question of what might happen with a more complex relationship between weights and states, or if more elaborate architectures were employed to approximate the value function. For example, a network with a hidden layer between a look-up table state representation and the value output might be used to capture underlying structure in the mapping from states to values. A related idea is that of a "basis function" representation of state. This is directly relevant to the hypothesis that place cells might provide a representation of state, since place cells can be considered as radial basis functions of the underlying location space.

There are two parallel motivations for making use of function approximation. On the one hand, it is a way of reducing the dimensionality of large MDPs, by approximating a large state space with fewer functions and so fewer parameters. On the other hand, it is a way of speeding up the transmission of credit information from temporally distant parts of the state

space, by specifying relationships between the values of separated states. Note, however, that relationships which hold for a particular optimal value function, *ie* for a particular transition function or set of returns, need not necessarily hold under a different transition function or a different set of returns. In this way, the aim of dealing effectively with large state spaces is not entirely equivalent to the aim of obtaining generalisation.

Sutton (1988) originally suggested using TD learning with a multi-layer perceptron architecture, using the back-propagation rule to train weights (*ie* to specify weight changes in layers other than the last). Back-propagation networks are usually used to perform non-linear regression, for which they can be very effective. They are often very poor, however, when applied as suggested to TD learning, and other forms of RL. A rather more restricted class of function approximation schemes appears to work in RL, and place cells as modeled belong to this class. First, with a mind to place cells, radial basis functions are introduced as a form of function approximation which, even in the context of supervised learning, is different in important ways from the multi-layer perceptron. Second, the performance in RL of different forms of function approximation is briefly reviewed, in terms of both empirical studies and theoretical results, both of which lend computational support to the use of place cells with TD learning.

3.5.1 Radial Basis Functions

Given a set of N data points $\{(\mathbf{x}_0, y_0), (\mathbf{x}_1, y_1), ...(\mathbf{x}_N, y_N)\}$ (where $(\mathbf{x}_i, y_i) \in R^n \times R$) from some unknown function $y(\mathbf{x})$, an interpolation can be made of the form (Powell, 1987):

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{N} w_i \phi(||\mathbf{x} - \mathbf{x}_i||) \tag{3.22}$$

where each of N *radial basis functions* (RBFs), $\phi(||\mathbf{x} - \mathbf{x}_i||)$, is some function of the Euclidean distance between the input variable, $\mathbf{x}$, and a data value, $\mathbf{x}_i$. With each RBF is associated a weighting w_i which must be determined. One example of a function that might be used for $\phi(||\mathbf{x} - \mathbf{x}_i||)$ is the gaussian function of equation 2.1 which was used to model the firing rate of a place cell.

Poggio and Girosi (1990) have investigated RBF methods from the perspective of *regularisation*. Regularisation is a way of controlling the smoothness of the approximating function. A quantitative form of Occam's razor, smoothness is a prior assumption which guards against over-fitting and so promotes generalisation. Poggio and Girosi consider the interpolative problem addressed in equation 3.22 as that of determining a function f that

minimises the functional

$$H[f] = \sum_{i=1}^{N}(y_i - f(\mathbf{x_i}))^2 + \lambda||Pf||^2 \tag{3.23}$$

where the second term is the regularisation term, P, a differential operator (with respect to $\mathbf{x}$). The specific form for P will depend on the particular problem being solved. For example, in certain cases where $f(\mathbf{x})$ is a linear function of a set of weights, an appropriate choice of operator leads to a regularisation term proportional to the sum of the squared values of the weights, a strategy alternatively referred to as ridge regression (Hoerl and Kennard, 1970) or, in the context of connectionist approximation architectures, weight decay (Hinton, 1987). Under certain contraints on P, and using a variational method, RBFs emerge as a solution to equation 3.23, *ie*

$$f(\mathbf{x}) = \sum_{i=1}^{N} w_i G(||\mathbf{x} - \mathbf{x}_i||) \tag{3.24}$$

where $G(||\mathbf{x} - \mathbf{x}_i||)$ is a radially symmetric *Green's function* centered at point $\mathbf{x}_i$, and the weights, w_i, are given by the linear calculation: $(\mathbf{G} + \lambda\mathbf{I})\mathbf{w} = \mathbf{y}$, where $\mathbf{G}$ is an $N \times N$ matrix of elements $G_{ij} = G(||\mathbf{x}_i - \mathbf{x}_j||)$, and $\mathbf{w}$ and $\mathbf{y}$ are vectors of length N, with elements w_j and y_i respectively. Although not all suitable forms of RBF are also suitable Green's functions (*eg* multi-quadric functions are not; Hardy, 1971; Micchelli, 1986; Broomhead and Lowe, 1988), gaussian basis functions, such as might be provided by place cells, are.

More generally, a function might be approximated using M RBFs, where $M << N$, and where each RBF has an associated "centre", $\mathbf{c}_i$, no longer related to any particular data point (Broomhead and Lowe, 1988):

$$f(\mathbf{x}) = \sum_{i=1}^{M} w_i \phi(||\mathbf{x} - \mathbf{c}_i||) \tag{3.25}$$

This form of approximation is clearly equivalent to the kind of connectionist network architecture discussed in section 3.4.1 (figure 3.1).

The usefulness of this approach is that a linear dependence on the variable weightings, w_i, is combined with an ability to model explicitly non-linear relationships, due to the non-linear form of the $\phi(||\mathbf{x} - \mathbf{c}_i||)$ functions. Following theoretical work by Micchelli (1986), Broomhead and Lowe (1988) demonstrated that, for supervised learning of functions of the type of equation 3.25, assuming suitable choices of centres, $\mathbf{c}_i$, and functions, $\phi(.)$, there will exist a uniquely determined set of weightings, w_i. This distinguishes RBFs from other non-linear approximation methods, such as multi-layer perceptrons (*eg* trained by back-propagation on supervised learning problems).

The analysis of Poggio and Girosi (1990) extends to the case of the more general function approximation scheme of equation 3.25, in which the number of RBFs is usually considerably smaller than the number of training data. However, Poggio and Girosi take the view that in this case, supervised learning should also be used to alter the parameters of the RBFs themselves. A suggestion closer in spirit to that considered in this thesis is described by Bishop (1995), that RBFs can be determined from a set of input data using unsupervised learning techniques, with subsequent learning of the weights, w_i, using fast linear fitting.

In summary, place cells as RBFs may provide powerful representations for learning, in a guaranteed and fast manner, non-linear functions of the input data (presumably itself highly processed sensory data; see chapter 2). They additionally lend themselves to regularisation techniques. Their suitability for RL methods, and TD learning in particular, is discussed in the next section.

3.5.2 Function Approximation Schemes and RL

The failure of function approximation in general to support value function learning has been demonstrated by Boyan and Moore (1995). They considered a variety of control learning tasks, to which they applied value iteration (section 3.3.3) in conjunction with a greedy controller, *ie* one which always chooses the optimal action as evaluated from immediate rewards and next-state values: (1) navigation of a simple, two-dimensional navigational space without obstacles; (2) navigation of an environment similar but for the addition of two localised regions of high cost that should be avoided; and (3) a "car on the hill" task in which both location and velocity are (one-dimensional) state variables, and the optimal value surface is discontinuous.

The first point they draw attention to is that even if the optimal value function is simple (*eg* linear for task 1 with the fixed cost per transition model they use), more complex sub-optimal value functions need to be represented *en route* to this optimal end-point, which cannot be represented by a low order polynomial. Hence they report that for task 1, a global, linear approximation (which reduces the learning problem to that of estimating 3 parameters) converges only "luckily", *ie* after a very large number of iterations, and that a global, quadratic approximation diverges away from the optimal solution.

The second and potentially more worrying point is that these infelicities also affect more flexible approximation schemes. A multilayer perceptron with one hidden layer was classified as "lucky" to converge on task 1, and diverged on tasks 2 and 3. A different approximation scheme known as local weighted regression (Cleveland and Devlin, 1988) while proving succesful on tasks 1 and 3, also diverged on task 2.

However, the pessimistic conclusions of this work have been disputed by Sutton (1996), who demonstrated convergence with function approximation on both task 2 and task 3 from Boyan and Moore (1995), but with two apparently critical differences with respect to the solution. First, an *online* learning algorithm was used, that is, the SARSA algorithm of Rummery and Niranjan (1994; also Singh and Sutton, 1996), which can be considered as an extension of Watkins' Q-learning analogous to TD(λ) as an extension of TD(0). Second, a different function approximation scheme was used, known as the CMAC representation (Albus, 1981; Miller *et al*, 1990; Watkins, 1989). The CMAC is a way of creating localised, distributed representations of multi-dimensional data which both reduce the number of parameters required to represent the data (*ie* provide a more efficient coding than that of a punctate, localist representation), and also provide an appropriate basis for a degree of generalisation (*ie* interpolation). As an example of a CMAC, a two-dimensional navigational state space of 9×9 states can be represented in a way which uniquely distinguishes each state by two coarse grids of 5×5 "tiles" which are independently active, and which are displaced relative to each other by one state in each direction (figure 3.4). The result of this distributed code is the use of fewer units ($2 \times 5 \times 5$ tiles); the greater the number of tiles, the greater the efficiency (and, less directly relevantly to navigation, the scheme also becomes progressively more efficient the greater the dimensionality of the space). However, the code also supports generalisation by effectively providing a stepped version of a localised basis function code, *ie* it is potentially more useful than a completely distributed code. In fact, the CMAC resembles the RBF approach, and indeed Sutton remarks that the results he has found for CMACs might be expected also to be found using RBFs.

It remains to note two theoretical results which provide concrete guarantees that value learning will converge using function approximation. Tsitsiklis and Van Roy (1996) consider a discrete space of states $\mathbf{x}_i, i \in 1, ..., N$, and in a slight twist to the interpolation vs. regression dichotomy presented so far, they consider that a subset of states are associated with an RBF, $f_k(||\mathbf{x} - \mathbf{x}_k||)$, where $k = \{1, ..., M\}$ and $M << N$. Provided that the spread of the RBFs used is made small enough that

$$\delta \equiv \max_{k \in \{1,...,M\}} \sum_{i \neq k} |f_k(\mathbf{x}_i)| < 1 \tag{3.26}$$

and that, for some $\gamma' \in [\gamma, 1)$, where γ is the discounting factor, for all states $\mathbf{x}_i$ in the data set,

$$\sum_{k=1}^{K} |f_k(\mathbf{x}_i)| \leq \frac{\gamma'}{\gamma}(1 - \delta) \tag{3.27}$$

then convergence is assured (within a bound which gets looser as γ' gets bigger) for a dynamic programming procedure which combines value iteration procedure with greedy

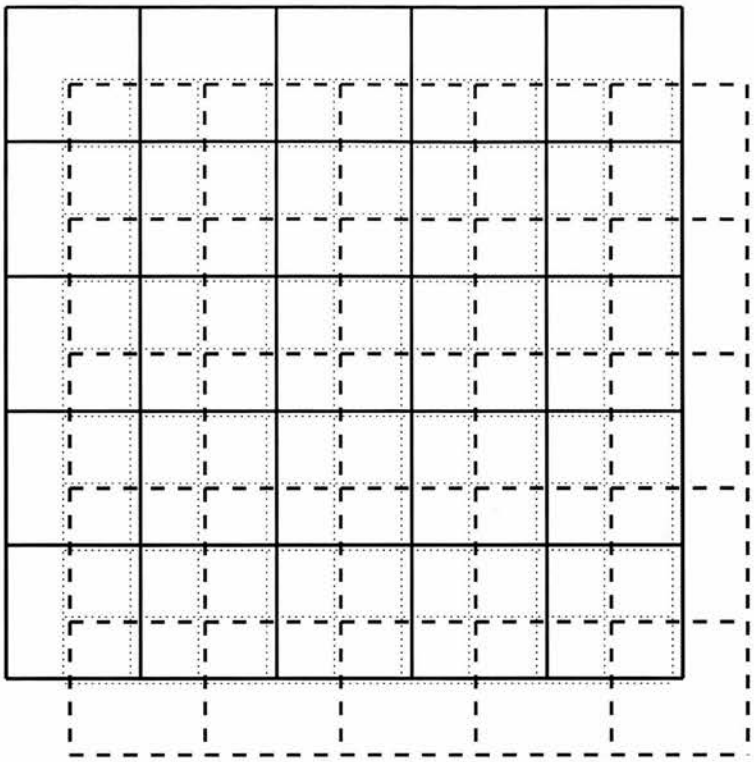


Figure 3.4: The CMAC principle: a 9×9 state space (shown with feint lines) can be represented in a distributed manner using two overlapping 5×5 grids (shown with thick lines, one solid and one dashed). Coarser grids may be used if in greater number, *ie* if the representation becomes more distributed.

control. However, Tsitsiklis and Van Roy (1996) note that, in practice, the spread of basis functions is likely to be set at a higher level than this theoretically safe value.

Perhaps of most relevance, however, is the theoretical *tour de force* of Tsitsiklis and Van Roy (1997), guaranteeing the convergence of TD(λ) learning with basis function approximation. They assume an underlying infinite-horizon, discounted Markov chain (with a possibly infinite state space) to which TD learning is applied (therefore not addressing the problem of optimising control, such as might use actor-critic learning). Under a number of standard, formal assumptions, they have proved that for any $\lambda \in [0, 1]$, the TD(λ) algorithm with a linear function approximation scheme (such as that of equation 3.25) converges with probability 1 to the unique solution allowed by the approximation scheme. Further, they establish a bound on the approximation error the tightness of which is governed by the factor $\frac{1-\lambda\gamma}{1-\gamma}$. An important insight afforded by the theory concerns the importance of online sampling, which ensures that state transitions are sampled with the frequencies natural to the Markov chain, and the fact that without such a guarantee divergence is possible, which may provide an explanation for some of the differences between the empirical results of Boyan and Moore (1995) and Sutton (1996) discussed above. However, Tsitsiklis and Van Roy also present an example of the divergence of online TD learning when using a function

approximator that is non-linear as a function of parameters, suggesting that the theoretical importance of RBF approximation (which although non-linear as a function of state, is linear as a function of parameters) will not necessarily diminish in the light of future convergence results.

It is worth noting that convergence results, even those as comprehensive as this, can be somewhat limited in what they reveal about the dynamics of an algorithm. For example, the above bound on the approximation error depends on λ in such a way that the error is best (*ie* smallest) for $\lambda = 1$, but deteriorates as λ falls, and is worst for $\lambda = 0$. Thus, the analysis fails to capture the importance during learning of the $\lambda < 1$ case; the theory is oblivious to the bootstrapping effect. Without more revealing analyses, simulation studies are necessary to investigate these kinds of effects.

3.6 Empirical Evidence

3.6.1 Animal Learning In General

Temporal difference learning has provided a simple but powerful model of associative learning in classical conditioning, effectively extending the Rescorla-Wagner rule to the temporal domain (Sutton and Barto, 1987). In particular, TD provides an explanation for second-order conditioning, whereby a conditioned stimulus, or CS, that has acquired predictive value, can itself condition another preceding CS.

Direct evidence for prediction at a neural level of the sort we are interested in has been reported. Dopamine neurons of the primate ventral tegmental area fire during conditioning tasks not to the presence of a reward but apparently to its expectation, such that after sufficient training in a classical conditioning task they come to fire to the conditioned cue rather than to the reward (Schultz *et al*, 1997). This is consistent with the TD prediction error (Schultz *et al*, 1997; Montague et al., 1996). On the basis of this evidence, Montague *et al* (1996) built an actor-critic model of rewarded conditioning behaviour in which neurons in the ventral tegmental area and the substantia nigra pars compacta report the prediction error term, and the dorsal striatum plays the role of the actor.

From a neural perspective, the job of constructing and using a temporal difference is arguably more biologically plausible than that of constructing a supervised learning error signal: (1) the temporal difference signal is obtainable as the first derivative of the dendritic sum of a value function approximation network such as that of figure 3.1, and neurally plausible ways of obtaining such derivatives are readily conceivable (*eg* cortical gain control through short-term synaptic depression; Abbott *et al*, 1997); (2) the requirement for

policy learning is that the prediction error should be (spatially but not temporally) diffusely available to modulate plasticity between state representation and actor representation, and with regard to the dopamine system discussed below, diffuse dopamine projections may indeed modulate long-lasting synaptic plasticity (Seidenbecher *et al*, 1997; Frey and Morris, 1997).

3.6.2 Navigation In Particular

The question remains whether such mechanisms are involved at all in navigational learning. A key issue is the extent to which there is any interaction between hippocampus and predictive learning areas such as those described above. Although the dominant theory of how the hippocampus contributes to navigation emphasises *independent* roles for the hippocampus (locale processing) and such areas as the striatum (route processing) in navigation (**Cognitive Map**; O'Keefe and Nadel, 1978), nevertheless it is possible (and hypothesised in this and the next chapter) that at least some aspects of navigational learning may depend on hippocampus and striatum working together. Clearly this is required if place cells are to provide a state representation which is then used for value and policy learning. The following findings are pertinent: (1) Lesions of the dorsal striatum (also known as caudate-putamen) result in a severe impairment on acquisition and retention of the RMW task (Whishaw *et al*, 1987; Devan *et al*, 1996); (2) Perhaps correspondingly, dorsal striatum lesions also impair acquisition of the reference memory component of a 12-arm radial arm maze task (Colombo *et al*, 1989); (3) Interestingly, hippocampally dependent tasks that appear immune to striatal damage are also those with no temporal credit assignment component, or where that component has been effectively removed, *eg* place-response learning in a T-maze (Oliveira *et al*, 1997), or learning in a watermaze to approach one of two visible platforms, given that only one is stable (the other flips over if the rat attempts to climb on) and where the correct one occupies a fixed spatial location (Packard and McGaugh, 1992).

Joint hippocampal and striatal learning may be supported by identified neural pathways. Both the ventral and dorsal striatum of the rat receive outputs from the CA1 hippocampal subfield, an area where place cells are found (Wiener, 1996). Electrophysiological data show location-specific unit responses in areas of the striatum related to the hippocampal system (Lavoie and Mizumori, 1994; Wiener, 1993).

In conclusion, while evidence concerning specific neural processes during navigation is scant, and certainly too much so to accept or reject outright a suggested learning scheme, nevertheless there is a small amount of suggestive evidence that mechanisms similar to those discussed in this chapter might contribute to some aspects of navigational learning.

3.7 Conclusion: MDPs and Navigation

At this point, a number of the aims set out in chapter 1 have been fulfilled. A principled, mathematical approach has been identified, in the form of MDPs, with which to characterise and investigate navigation as a problem. The approach arguably highlights rather than disguises many of the computational difficulties inherent in navigation. Moreover, a number of rather simple and apparently neurally plausible algorithms are available for solving problems of this sort. Finally, the stage has been set for an interpretation of the role of hippocampal place cells in navigation which, it was argued in chapter 2, has hitherto presented something of a paradox.

A number of simplifications have been made. For example, time has been treated as a discrete quantity. However, algorithms such as TD learning extend straightforwardly to the case of continuous time, essentially only requiring that reward and the discount factor be specified as functions of transition duration (Bradtke and Duff, 1995). The second assumption that was initially made was of a finite state space, but clearly function approximation schemes (such as RBF approximation) can solve this problem by approximating a continuous state space with a finite set of parameters. The most important, implicit simplification is the treatment of navigation as a problem of essentially fixed rewards – whereas reward structure (or even transition structure) can change, as when the location of the goal changes in the DMP task. The remainder of this thesis is concerned, therefore, with the problem of learning to navigate to multiple goals within a familiar environment.

One way of dealing with generalisation in MDPs might be to learn the complete reward and transition functions. For example, goal generalisation might then emerge by combining an existing transition function with a new reward function. The argument against model learning was given at the beginning of section 3.4, as the fact that it requires learning many more parameters. However, the extra effort might perhaps be justified if systematic changes in goal position were really expected.

Imagine, then, a simple supervised learning algorithm for attempting to estimate next-state probabilities as a result of various actions. Clearly, the current policy will have a dominant effect on the estimates that result. One simplification might be to learn something about the next-state probabilities under a random policy. However, this is a significant imposition on the exploration behaviour of the animal or agent. The problem is demonstrated by a related algorithm, Dayan's successor representation scheme (SR; Dayan, 1993). The SR attempts to learn something like a decomposition of the value function into reward and transition components. The problem is that as a single policy is pursued, so the transition component that is learned comes to reflect only that particular policy (*ie* that particular goal location),

and so generalisation to a new goal location is no easier. A period of "latent learning" (unrewarded random exploration) can facilitate subsequent learning, but for animals goal generalisation is clearly possible without this kind of exploration.

Therefore, even in multiple goals MDPs, learning a complete model of the transition function may not be a viable strategy. Instead, a basis for generalisation is sought, which may take a number of forms. In the following chapter, a model-free method is presented for learning a geometrical representation of an environment, which can support one-trial learning to novel goal positions in a limited class of environments, but which can be learned quickly, and to a large extent independently of the current policy being pursued by the animal. In chapter 6 a more general set of techniques are investigated for learning the underlying structure of environments, which can be made use of by model-free methods in multiple goals problems.

Chapter 4

A Hippocampal Model of One-Trial Spatial Learning Using Temporal Difference Learning

4.1 Introduction

I have chosen to model the RMW and DMP watermaze tasks described in chapters 1 and 2, as both are highly sensitive to hippocampal lesions and representative of the kind of navigational problems for which TD learning, used in conjunction with place cells, might provide a solution. To recap, the acquisition rates of rats on these tasks are shown again in figure 4.1a (for RMW, days 1 to 7, and reversal to a novel platform position on days 8 and 9) and figure 4.1b (for DMP). The key features of this learning which it is hoped to capture are the rapid acquisition of RMW performance, the almost one-trial learning evident in the reversal, and, for DMP, the gradual change from gradual learning on the first day to one-trial learning to a novel platform position by the sixth.

RMW has been modelled as an instance of conventional reward-based learning using place cells (Brown and Sharp, 1995; see also Burgess *et al*, 1994). However, the task presents a distal reward problem, which these models do not really solve. It has been dealt with in different ways – *eg* by postulating very large place fields covering the entire environment, although these are rarely observed (Burgess *et al*, 1994), or by making use of a memory trace, for which there is no evidence over the kinds of distances required, and which in any case leads to a rather inefficient learning algorithm (Brown and Sharp, 1995). A place cell very far from the goal can learn an expectation of reward simply by maintaining a trace memory of its activation which decays so slowly that when the animal gets to the goal, a residual trace will remain. However, this is inefficient because an animal's paths will be extremely variable during learning – early on in training, the animal sometimes

b

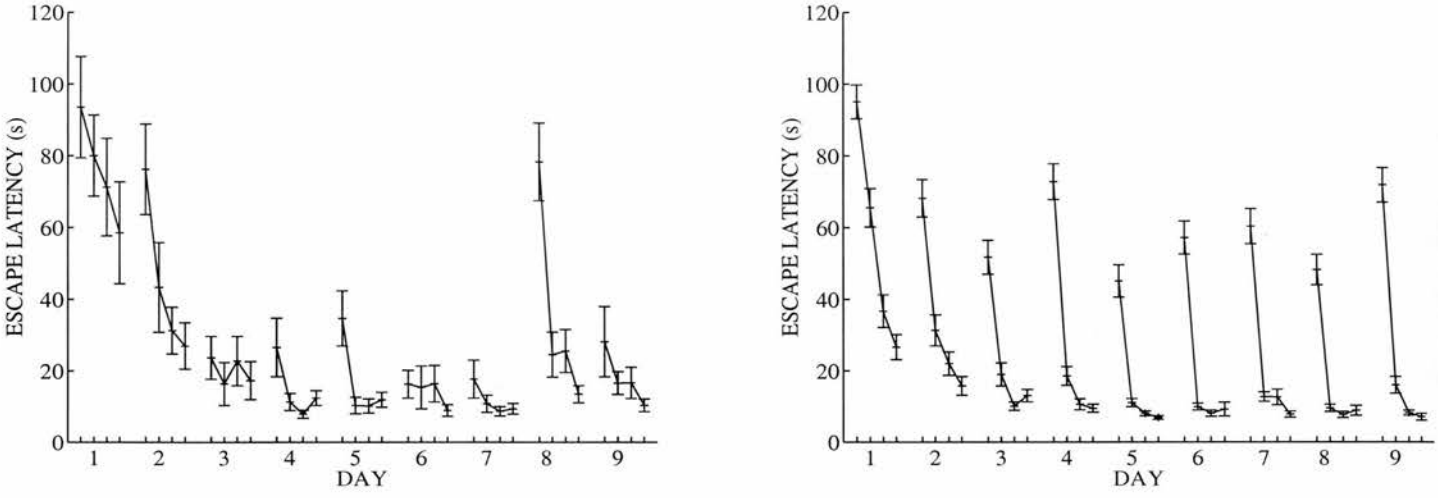


Figure 4.1: The performance of rats on (a) reference memory (RMW), N=12, and (b) delayed matching-to-place (DMP), N=62. In both figures, escape latency (time taken to reach the platform) is plotted across days (RMW task: 4 trials/day, fixed platform location days 1-7; reversal to new platform location, days 8-9; DMP task: 4 trials/day, new platform location each day). Note (i) asymptotic performance in RMW task, (ii) one-trial learning in DMP task, (iii) difference in escape latency on second trial of day 8, between the two tasks. Trial 1 performance differs from day to day, due to the platform position, which was the same for all rats on any given day. It was observed that platforms nearer the centre of the pool, or near to a starting position, were easier to find under random search than others. Figure 1b is from Steele and Morris (1999); the data for Figure 1a were obtained in the same apparatus and using the same methods as those described for the DMP task by Steele and Morris (1999), excepting that: (1) the platform remained in the same location across days, until moved to the opposite quadrant on Day 8; and (2) the intertrial interval was always 15 sec.

gets to the goal quickly and sometimes not – and the residual value of a particular place cell's trace will likewise be extremely variable from trial to trial. Unfortunately, it is these residual values that Brown and Sharp's learning rule must average over. The hope is that TD learning provides a more efficient method, essentially because the the difference between successive estimates of value may be in general less variable than the whole path lengths that Brown and Sharp's algorithm effectively considers (bootstrapping in TD learning; *eg* see section 3.4.5).

Therefore a simple "actor-critic" model of RMW learning is first considered (Barto *et al*, 1983; Barto *et al*, 1990), in which a set of place cells is associated with a representation of reward expectation, and also with a representation of action choice. Critically, the TD learning rule is used to predict rewards.

DMP is computationally more demanding than RMW. Unlike RMW, this task involves al-

tering actions after only one trial of experience. It does not, however, only involve rapid learning, as is demanded in a standard delayed match-to-sample task. DMP in the water-maze is a complex navigation task in which a whole sequence of navigational actions has to be inferred from the single learning experience. This suggests that rats learn a representation of space that is goal-independent, which, following many other models, we model as a metric coordinate system, learned from self-motion information (Wan *et al*, 1993; Redish and Touretzky, 1995; also Blum and Abbott, 1996; Gerstner and Abbott, 1997). However, these attempts at modelling coordinate learning have encountered a global consistency problem. While a natural basis for learning the coordinates in the first place is the self-motion (or "dead reckoning") information which an animal has available, such information, while suitably metric, is only *relative* in nature. Simply performing path integration on this information runs into trouble as soon as the animal loses track of its origin – as must happen during laboratory navigation tasks in which an animal is often picked up from the goal location at the end of one trial, and started again from an unpredictable starting location. If the animal path-integrates from each new starting position, it will quickly acquire inconsistent coordinates over the environment as a whole.

Therefore, in this chapter, a novel application of TD learning is investigated that develops consistent coordinates directly, but which nevertheless uses hippocampal place cells in the same way as reward learning – as a stable representation of state.

The chapter begins by presenting the reward-based component of the model, demonstrating that this component alone captures some aspects of spatial learning, but not all. In particular, it does not capture the flexible way in which rats can learn about novel goal locations. The second component of the model, the learned coordinate system, is then described, along with a simple way in which the components can be made to work together. Simulation results are presented which capture performance in both RMW and DMP tasks. The discussion addresses the role of place cells within the model, what can be inferred from the model about the nature of the two tasks, and the relationship of the model to experimental data and to other models of hippocampal function. Finally, a set of novel experimental predictions is presented.

4.2 Reward-Based Navigation

Consider a simulated animal in an environment with control of its own actions. At any given time t, the animal is able to choose an action. Also at any given time t, the environment provides the animal with a reward r_t. If the animal moves onto the platform (a certain region of the environment) at time t, $r_t = 1$; otherwise $r_t = 0$. The difficult problem is to

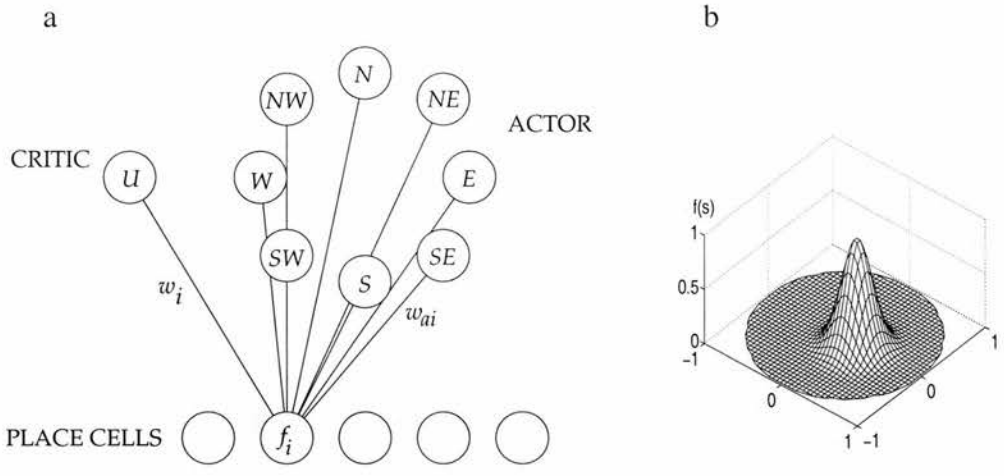


Figure 4.2: The actor-critic system. (a) An input layer of place cells projects to the critic cell, whose output, U, is used to evaluate behaviour. Each place cell also projects to 8 action cells, which the actor uses to select between 8 possible directions of movement from any given location. (b) An example of a gaussian place field (x and y axes represent location, z axis represents firing rate).

learn correct actions given such a sparse reward signal.

To solve this problem an actor-critic architecture was used. The implementation of the actor-critic has three parts (figure 4.2a): i) an input layer of *place cells*, ii) a *critic* network that learns appropriate weights from the place cells to enable it to output information about the value of particular locations, and iii) an *actor* network that learns appropriate weights from the place cells which enable it to represent the direction in which the rat should swim at particular locations.

Hippocampal Place Cells

As described in chapter 2, the activities of place cells are modelled as Gaussian functions of location in the maze (figure 4.2b). If the rat is at position $\mathbf{s}$, then the activity of place cell i ($1 \leq i \leq N$) is given by:

$$f_i(\mathbf{s}) = \exp\left(-\frac{\|\mathbf{s} - \mathbf{c}_i\|^2}{2\sigma^2}\right) \tag{4.1}$$

where $\mathbf{c}_i$ is the location in space of the centre of cell i's place field, and σ is the breadth of the field, equivalent to the radius of the circular contour where firing is 61% of the maximal firing rate. We consider an ensemble of place cells ($N = 493$) with place fields distributed in overlapping manner throughout the maze, each with width $\sigma = 0.16m$.

Although clearly idealised, these place cells illustrate the limitations pointed out in chapter 2 – they are not intrinsically informative about spatial or navigational quantities such as dis-

tance or direction from a distant goal. However, such units form a radial basis function representation of location (section 3.5.1). As such they would support the representation and learning of functions which vary (usually smoothly) with location. This chapter explores this hypothesis – that hippocampal place cells play the limited but nonetheless critical role of providing a particular representational substrate.

The Critic

The critic constitutes a single unit, whose firing rate at a location $\mathbf{s}$ is given by a weighted sum of the firing rates of place cell inputs $f_i(\mathbf{s})$:

$$U(\mathbf{s}) = \sum_i w_i f_i(\mathbf{s}). \tag{4.2}$$

where w_i is the weight from place cell i. At each time step, a prediction error, δ_t, was calculated, using equation 3.7 (with $\gamma = .9975$). There was one exception to this rule: when the rat was *on* the platform (*ie* when $\mathbf{s}_t$ was within the goal area), the prediction error was instead given by:

$$\delta_t = r_t - U(\mathbf{s}_t) \tag{4.3}$$

This is an amendment to the conventional TD rule, but it makes sense for absorbing Markov decision problems such as this, for which the value of the absorbing state (*ie* platform location) is unspecified. Note that in the case of a look-up table representation of state, the issue is of less practical importance, since the value of an absorbing state will remain as initially set. However, for the case of function approximation (as here), the absorbing state *is* likely to change in value. It was found that incorporating this amendment increased considerably the highest learning rate at which stable convergence could be achieved.

The critic weights, w_i, were changed using:

$$\Delta w_i = \eta \delta_t f_i(\mathbf{s}_t) \tag{4.4}$$

ie TD(0). Following standard reinforcement learning practice, a fixed learning rate was used to avoid slow learning. The price to be paid is residual error. However, the results suggest that this error is insignificant.

The Actor

For convenience, the rat is allowed to move in one of eight possible directions at each time step (north, northeast, east etc.) represented by eight action cells $a = 1 \ldots 8$. The activity

of each action cell at position $\mathbf{s}$ is:

$$\rho_a(\mathbf{s}) = \sum_i w_{ai} f_i(\mathbf{s})$$

where w_{ai} is the weight from place cell i to action cell a. The swimming direction is then chosen stochastically with probabilities $p(\mathbf{s}, a)$ related to these activities by:

$$p(\mathbf{s}, a) = \frac{e^{2\rho_a(\mathbf{s})}}{\sum_{a'} e^{2\rho_{a'}(\mathbf{s})}} \tag{4.5}$$

Actor weights w_{ai} were then changed using:

$$\Delta w_{ai} = \eta^A \delta_t f_i(\mathbf{s}_t) g_a(t), \tag{4.6}$$

where δ_t is calculated as for the critic, $g_a(t) = 1$ if action a was chosen at time t, and $g_a(t) = 0$ otherwise.

4.2.1 Performance Of Reward-Based Navigation

Simulation procedures

The swimming behaviour of a rat was simulated in a 2m diameter circular watermaze, which contained a 0.1m diameter escape platform. These parameters are the same as those in (Steele and Morris, 1999). The swimming speed of the rat was constant at 0.3ms^{-1}. The walls were treated as reflecting boundaries – the rat 'bounced' off. Any move into the platform area was counted as a move onto the platform. Space was treated as a continuous variable, however, time was discretised into steps of 0.1s. Simulations with 0.01s bins produced similar results to those with the coarser discretisation, and so show that this discretisation does not produce artefacts.

In reality, a rat cannot choose a different direction at the fine-grained time steps of the temporally discrete simulation. To model momentum, the direction the rat heads was given by a mixture of control as specified by the actor, and the previous heading, in the ratio $1 : 3$. This restricts the turning curve of the rat, and is particularly important early on, when the whole pool must be searched fairly quickly. One technical concern about momentum is that it means that the path to the goal from a location is partly determined by the direction in which it was swimming when it arrived at that location. This disturbs the formal theory, although simulations demonstrate that it does not prevent good performance by the simulated rats.

Following the experimental protocols, each trial began at one of four starting locations located at the north, south, east and west edges of the pool, and ended when either the rat reached the platform, or a time-out of 120s was reached. For RMW, the platform remained

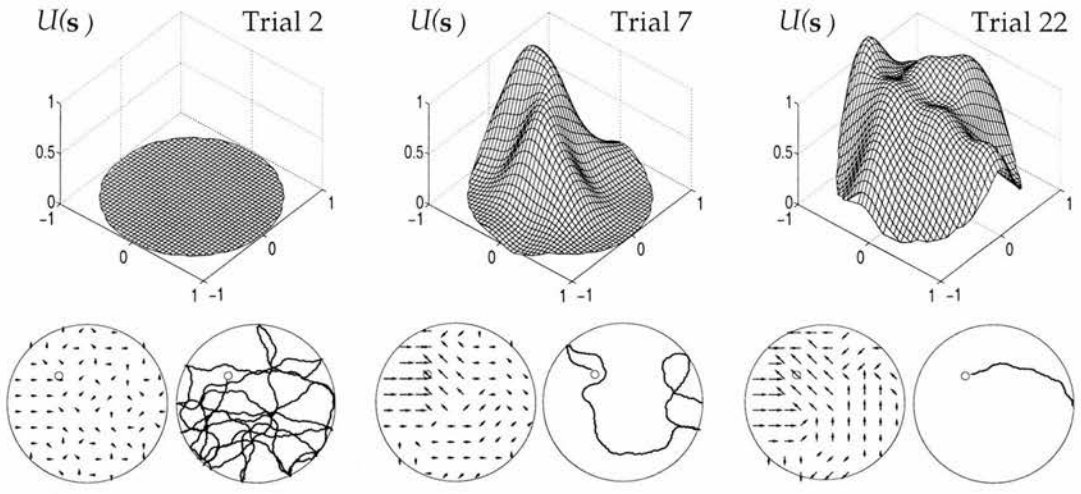


Figure 4.3: Learning in the actor-critic system in RMW. For each trial, the value function $U(s)$ is shown in the upper, three dimensional plot; to the lower left, the preferred actions at various locations are shown (the length of each arrow is related to the probability that the particular action shown is taken by a logarithmic scale); to the lower right is a sample path. Trial 2: after a timed-out first trial, the value function remains zero everywhere, the actions point randomly in different directions, and a long and tortuous path is taken to the platform. Trial 7: the value function being peaked in the north-east quadrant of the pool, the preferred actions are correct for locations close to the platform, but not for locations further away. Trial 22: the value function has spread across the whole pool and the preferred actions are close to correct in most locations, and so the actor takes a direct route to the platform.

in the same location throughout the simulation. In DMP, the platform was moved to a novel location after every four trials.

The learning rate parameters, which determine the constants of proportionality in equations 4.4 and 4.6, were optimised to give $\eta = .15$ and $\eta^A = .45$.

Simulation results

Figure 4.3 shows the gradual development of the value function. For the first few trials, it is informative about only a small area close to the platform location. Later in learning, however, values have spread out to all parts of the environment. This enables appropriate actions to be learned, as is reflected in ever shorter paths to the platform.

The actor-critic model of figure 4.2 was first applied to the reference memory (RMW) task. Figure 4.4a shows that the actor-critic captures learning in this task; path lengths reach asymptotically low values as quickly as the latencies of rats shown in figure 4.1a. However, when the platform is moved during the reversal phase of days 8 and 9, this model diverges from the performance of rats.

Likewise, when applied to the delayed matching-to-place (DMP) task, the results are strik-

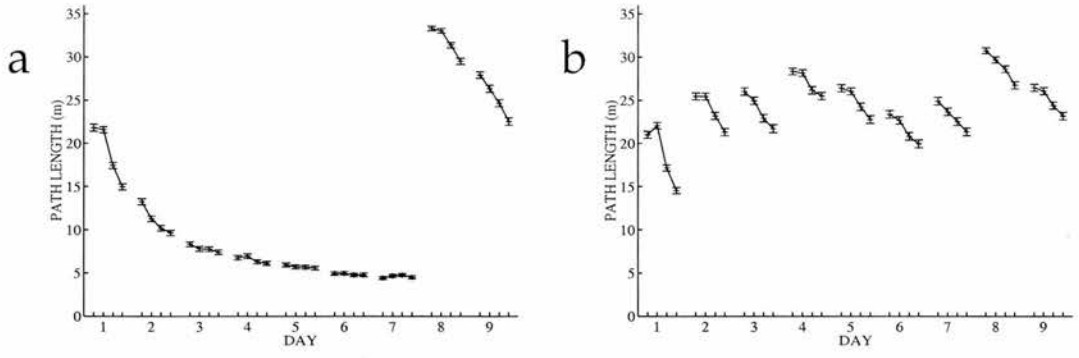


Figure 4.4: Performance of the actor-critic model. For each data point, the mean and standard error in the mean are obtained from 1000 simulation runs. (a) RMW task, in which the platform occupies the same location. The actor-critic captures acquisition, producing direct paths after around 10 trials. For the last eight trials however (days 8 and 9), the platform is moved to a different position (reversal), and the model fails to adapt rapidly enough. These simulation results can be compared to figure 4.1a. (b) DMP task, in which the platform remains in the same position within a day, but occupies a novel position on each new day. The actor-critic model captures acquisition for the four trials of day 1, for which the task is indistinguishable from RMW. However, as a soon as the platform is moved, the actor-critic not only fails to generalise to the new goal location, but suffers from interference from the previous days' goal locations. Rats suffer neither of these limitations (figure 4.1b).

ingly different. Figure 4.4b demonstrates that the actor-critic component of the model fails by itself to capture the performance of rats in DMP, because the value function that is learned confounds spatial and reward information, and so neither the value function nor the policy are flexible to changes in reward location. The model incorrectly predicts that learning a new platform position is much slower because of interference from previous days.

4.3 Coordinate-Based Navigation

4.3.1 Learning Globally Consistent Coordinates From Self-Motion Information

The actor-critic is a general solution to the problem of navigating to a fixed goal location. Nothing is assumed about the shape or topology of the environment, and short paths to the goal would ultimately be learned even in the presence of complicated barriers. However, the actor-critic model fails by itself to capture the performance of rats in DMP for two reasons. First, it incorrectly predicts that learning a new platform position is much slower because of interference from previous days. Second, it provides no mechanism by which the experience of previous days can provide any help with learning a new platform position.

One trial learning by rats on DMP reveals that rats suffer neither of these limitations. Under

appropriate training conditions, rats can not only avoid interference between training on successive days, but can also generalise from experience on early days to help performance on later days. For example, the starting position on trial 2 of day 6 of training (figure 4.1b) may be in an area of the environment not explored on trial 1 of that day; nevertheless the rat swims immediately to the platform. Clearly, knowledge from previous days is being used.

The present model of coordinate learning is based around the the observation that the computations involved in the *dead-reckoning* abilities of animals could subserve an all-to-all navigation system for open spaces like a watermaze, if only the dead-reckoning coordinates could be made to be consistent across separate trials, *ie* tied to an allocentric representation of the environment. In effect, the problem is considered of making a dead reckoning system hippocampal dependent, that is, dependent on input from the place cell system, to account for one-trial learning in the DMP task.

Dead reckoning abilities have been documented in (at least) ants, bees, wasps, geese, gerbils, pigeons, rats and humans (Gallistel, 1990). These abilities are based on the availability of instantaneous estimates of the animal's self-motion, which can be integrated in order to calculate the direction back to a starting point. The availability of this information can, however, be dissociated from using path integration over these quantities to estimate position, and there are good reasons for doing so.

It is hard to acquire an appropriate coordinate system using path integration information alone because of the problem of consistency. When the rat is put in the maze in a new place, there is no way of ensuring that the dead reckoning coordinates it assigns are automatically consistent with those it has assigned elsewhere in previous traversals of the maze. The essential task for the model is learning this consistency (see also Wan *et al*, 1994). The key observation is that for every move that the rat makes, the difference between its estimates of coordinates at the ending and starting locations should be exactly the relative self-motion during the move. This consistency condition can be used as the basis for a TD learning rule for learning coordinates.

An important assumption is that self-motion estimates are defined allocentrically, that is, consistently across trials with respect to the environment. Although the problem of having a consistent report of head direction is quite similar to that of having consistent coordinates, there are reasons for taking the former as given. First, individual distal cues can readily approximate a stable, global compass direction. Second, there is experimental evidence for the very rapid establishment of highly stable compass-like representations (chapter 2, section 2.3.5). Vestibular disorientation can disrupt the stability of head direction cells (Knierim *et al*, 1995), but in the experiments being modeled, such manipulations were not used.

Figure 4.5 shows a simple model of learning and using coordinates. The coordinate system consists of two networks, one which learns X coordinates (as $X(\mathbf{s}) = \sum_i w_i^X f_i(\mathbf{s})$), and one which learns Y coordinates (as $Y(\mathbf{s}) = \sum_i w_i^Y f_i(\mathbf{s})$), both using inputs from place cells which act in exactly the same way as in the actor-critic model, each producing a firing rate $f_i(\mathbf{s})$ as a function of location $\mathbf{s}$. The choice of X and Y coordinates, or even just two orthogonal directions, is of course arbitrary – but the basic problem of making coordinates consistent will exist whatever particular coordinate system is used. The X and Y coordinates have been chosen for simplicity, and to illustrate clearly the learning problem.

As the rat moves around, the weights $\{w_i^X\}$ and $\{w_i^Y\}, i = 1, ..., N$ that define the coordinates are adjusted according to:

$$\Delta w_i^X = \eta^X \left(-\Delta x_t + X(\mathbf{s}_{t+1}) - X(\mathbf{s}_t)\right) \sum_{k=1}^{t} \lambda^{t-k} f_i(\mathbf{s}_k) \tag{4.7}$$

$$\Delta w_i^Y = \eta^Y \left(-\Delta y_t + Y(\mathbf{s}_{t+1}) - Y(\mathbf{s}_t)\right) \sum_{k=1}^{t} \lambda^{t-k} f_i(\mathbf{s}_k) \tag{4.8}$$

where Δx_t and Δy_t are the self-motion estimates in the x direction and y direction, respectively. Note the use of the full TD(λ) rule. For coordinate learning, variance in the estimates was expected to be lower than for the value learning in the critic, suggesting that a high value of λ would make learning fastest (chapter 3, section 3.4.5). Simulations confirmed this, with $\lambda = 0.9$.

4.3.2 Why TD-Based Coordinate Learning Is A Good Idea

As described in chapter 3 (sections 3.4 and 3.7), perhaps the most serious drawbacks to the strategy of learning a model of transition probabilities in an environment involve the constraints placed on the exploration strategies of animals. To correctly estimate even such a simple quantity as the adjacency of two states (*eg* two place fields) requires an even sampling of state transitions, or otherwise an explicit representation of policy (and thus the learning of many more parameters). A clear example of where policy intrudes upon model learning is the successor representation, which learns distributed representations based on predictions of transitions from one state to another – but as a particular policy is learned, so the representations become increasingly tuned just to that policy, and the scope for generalisation to new goal positions is reduced (Dayan, 1993). To support one-trial learning in DMP, a completely goal-independent model must be learned, and yet from the very first trial the rat must also be improving its actions with respect to the current platform location, *ie* a latent learning-like period of random sampling of transitions is simply at odds with the data to be modeled.

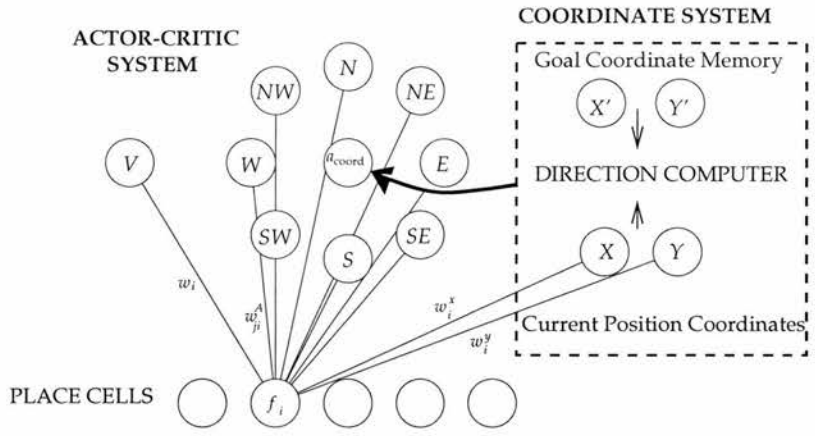


Figure 4.5: The combined coordinate and actor-critic model incorporates both the actor-critic system and a coordinate system. The coordinate system consists of three components: i) a coordinate representation of current position made up of two cells X and Y, the firing of which is a function of place cell input; ii) a goal coordinate memory consisting of two cells, X' and Y', whose firing reflects the coordinate location of the last place at which the platform was found; iii) a mechanism which computes the direction in which to swim to get from the current position to the goal. The output direction from the coordinate system is integrated with that from the actor-critic through the 'abstract action', marked a_{coord} which receives reinforcement depending on its performance.

It is therefore of considerable importance that TD-based coordinate learning is essentially insensitive to exploration strategy. First, the problem is purely one of prediction rather than control. Second, and unlike the case of, for example, the successor representation, the predictions are not sensitive to goal position, *ie* the constituent elements of the prediction error in equations 4.7 and 4.8 are independent of whatever the animal is attempting ultimately to do, provided the self-motion estimates are themselves correct (or variable according to a simple noise model). For this reason, it can be hoped that coordinate learning proceeds quickly and efficiently while a rat is trying to optimise its control to one goal in particular.

4.3.3 Using Coordinates To Control Actions

In dead reckoning, an animal computes, from its current coordinate, a bearing back to a point of origin. In the model, a coordinate controller computes, given its current allocentrically defined coordinate, a bearing to whatever other coordinate is of interest. This requires performing a simple vector subtraction, which is just the same computation that dead reckoning also requires (although the computation is not modeled explicitly in neural or connectionist terms). The additional, non-trivial requirement for the general coordinate system is some form of goal coordinate memory, a point returned to in the discussion. At certain times, however, there will be no remembered goal coordinate – during the first trial, and, on DMP, at every time the rat reaches the position where it thinks the goal is, and finds it to be moved. When there is no goal coordinate in memory, the coordinate controller

specifies random, exploratory actions.

When coordinates have been learned, a coordinate controller such as that described above is potentially extremely useful; however, if coordinates are poorly learned, there are no guarantees that the controller is at all useful. Early on, the controller will produce paths which are not only indirect, but are even prone to catastrophic loops (see figure 4.7). The ability of the controller to switch to random exploration can sometimes alleviate this problem, but even then is guaranteed to produce highly sub-optimal paths.

The solution adopted is to combine coordinate control with the actor-critic architecture. One way to do this is shown in figure 4.5. Here, there is an additional action cell, a_{coord}, representing the rat's preference for the swimming direction offered by the coordinate system. This coordinate action can be chosen stochastically, in competition with the normal actions, in a manner reminiscent of, but not quite similar to, Singh's (1992b) 'abstract actions'. The coordinate action is reinforced by the critic in exactly the same way as the other actions: when the coordinate action is chosen, the weighting of the coordinate action cell is changed according to current information from the critic. The only difference is that when there is no remembered goal coordinate – and the controller is specifying random exploratory actions instead of actions based on its coordinates – then the controller does not participate in learning *ie* a_{coord} is not updated. The effect is that coordinate control comes to be relied upon gradually, as it gives increasingly accurate information about where both the animal and the goal are located. Moreover, in a task for which coordinate control is irrelevant, it would be successfully ignored. Note that the coordinate system suggests appropriate actions without suggesting values associated with these actions.

4.3.4 Performance Of The Combined Coordinate and Actor-Critic Model

Simulation methods

The combined model was tested in simulated versions of the RMW and DMP tasks, using the same simulation environment as described for the actor-critic model. Learning rate parameters for the actor-critic were as in section 4.2.1. For equations 4.7 and 4.8, $\eta^X = \eta^Y = .01$ was used. For the learning rate governing choice of the "coordinate action", η^C, a choice was desired that would make learning comparable, in terms of action choice, to that of learning for the other actions, which were receiving a greater fan-in. A back of the envelope calculation (using equations 3.16 and 3.17) demonstrates that, after a weight update, the probability of a taking an action would be affected thus:

$$\begin{aligned} p(\mathbf{s}, a) &\approx \text{const} \cdot e^{\sum_i [w_{ai} + \eta^A \delta_t f_i(\mathbf{s})] f_i(\mathbf{s})} \\ &= \text{const} \cdot e^{\sum_i w_i f_i(\mathbf{s}) + \eta^A \delta_t \sum_i fi(\mathbf{s})^2} \end{aligned}$$

An optimised choice was found to be $\eta^C = 3.2$, which is close to that predicted by the above calculation, given that $\sum_i f_i(\mathbf{s})^2$ varied during simulated trials between approximately 8.5 and 8.7.

Simulation results

Figure 4.6a shows the development of the X and Y coordinates over days. Early on, *eg* day 2 trial 2, the coordinate surface is uneven. By day 6, it is relatively smooth. Note that the coordinate learning system receives no direct information about how the coordinates should be centred. Three factors control the centering: the boundary of the arena, the prior setting of the coordinate weights (in this case all were zero) and the position and prior value of any absorbing area (in this case the platform). These factors are arbitrary, and one might worry that the coordinates could drift over time and thereby invalidate coordinates that have been remembered over long periods. Consider, for example, a rat that had learned coordinates throughout a maze but was then confined for a period of time to a particular region of the maze. If the rat was later released, but coordinates had drifted in the meantime, navigation within the maze as a whole would be affected. However, since the expected value of the prediction error at time steps should be zero for any self-consistent coordinate mapping, such a mapping should remain stable. This is demonstrated for a single run: figures 4.6c and d show the mean value of coordinates $\bar{X}$ evolving over trials, indicating that there is little drift after the first few trials.

The difficulty in using the coordinates by themselves to specify actions is clear from the nature of the gradient of these functions (figure 4.7). Early on in learning, the coordinate functions are highly irregular, and a direction specified on the basis of these functions is worse than simply sub-optimal, since catastrophic loops are possible. This difficulty motivates the combination of the coordinate control with the actor-critic, allowing the conventional actions of the actor-critic to dominate early on, but enabling coordinate control to come to dominate as its actions prove more reliable than the conventional ones. This transfer of control happens rapidly during the DMP task (figure 4.6e).

Figure 4.8a shows the performance of the combined model in the RMW task. Like the actor-critic model of the previous section, the combined coordinate and actor-critic model successfully captures the acquisition of this task. Moreover, this model can also account for the rapid learning to the novel platform during the reversal phase, as seen in figure 4.1a. Figure 4.8b shows the performance of the combined model in the DMP task. Just as in figure 4.1b, acquisition during early days is gradual, while by day 6, one-trial learning is evident in the difference in performance between trials 1 and 2.

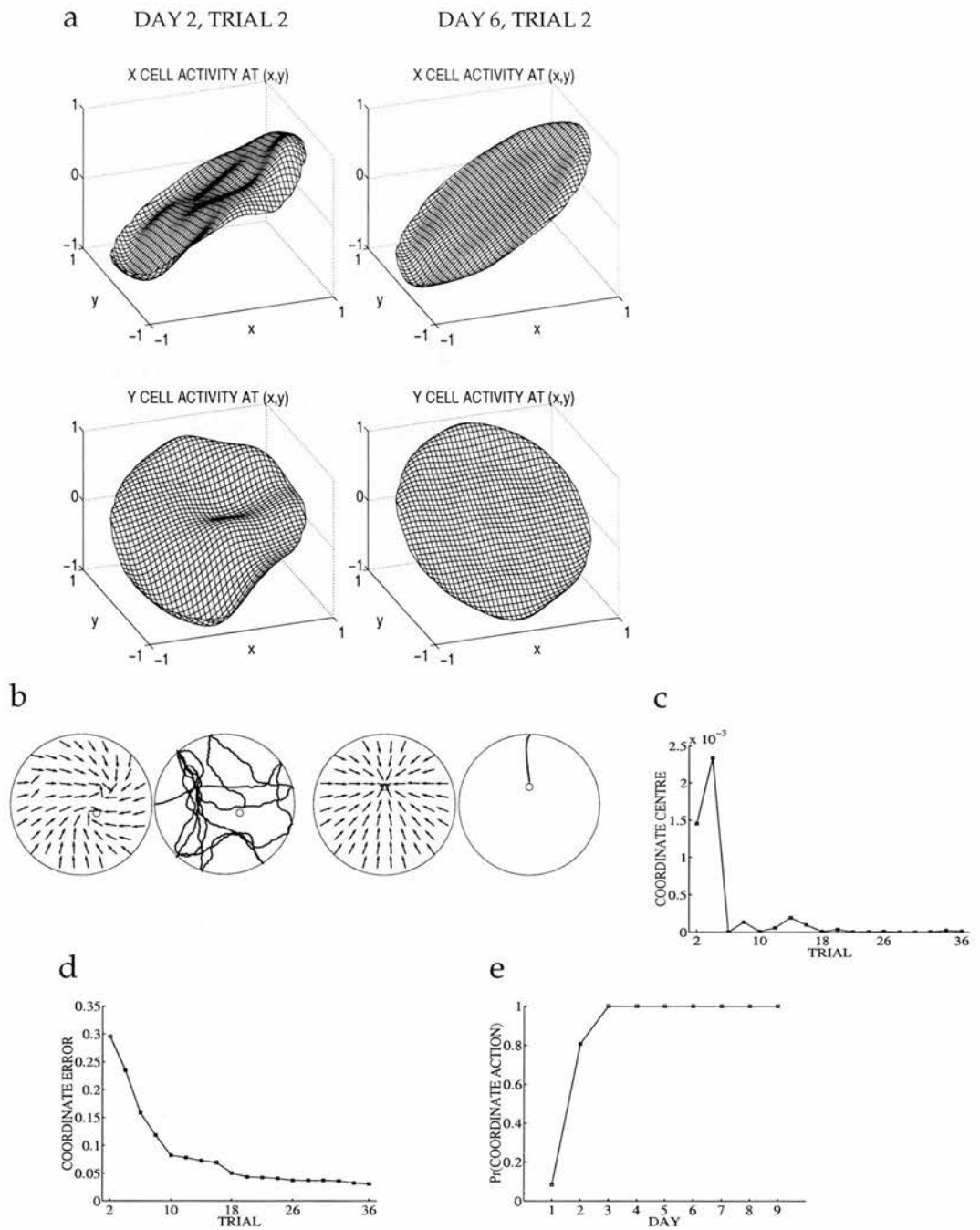


Figure 4.6: (a.) The X and Y coordinate functions develop gradually over days, at first being quite uneven (eg day 2) but becoming quite smooth by day 6. (b.) Below each coordinate figure are examples of preferred actions, and paths, for trial 2 of a simulated run of DMP using the full model. On the second trial of day 2, performance is quite poor. By day 6, one-trial learning is evident. (c.) The centering of the X coordinates, as measured by the mean, does not drift by the time coordinates are smooth. This is expected, since as the coordinates become consistent, all weight changes tend to zero. (d.) The error in the X coordinates for the same simulation, measured as the variance for each coordinate about its desired value relative to the mean. The error stabilises after a few trials. (e.) As coordinates improve, the weighting of the coordinate-based action increases. Thus the probability of taking the coordinate action, averaged over all time points within a trial, and over all the trials of a day, is shown to increase.

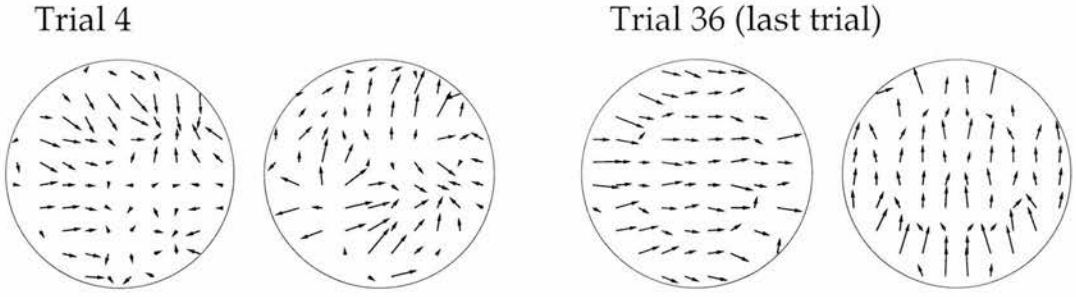


Figure 4.7: The gradient of the coordinate functions. The gradient is a very sensitive measure of smoothness. On trial 4, coordinates are still not at all smooth; navigation based on these functions alone would be prone to catastrophic loops, *ie* would never reach the platform. By comparison, the actor-critic scheme develops effective values and actions for control by trial 4 (figure 4.3), and it is this control that allows the rat to move through the environment, and so improve its coordinate functions. By trial 36, coordinates are smoother and the gradients reflect the X and Y directions.

4.4 Discussion

A model of hippocampally dependent navigation has been presented that uses place cells as a representational substrate for learning three different functions of position in an environment. The actor-critic component of the model learns the temporal proximity of locations to a single escape platform and also appropriate actions that get there quickly. By itself, the actor-critic model captures initial acquisition performance in RMW. However, its performance diverges from that of rats the moment the platform is moved, failing to account for the good reversal performance shown by rats, or for the even more striking one-trial learning in DMP. A second component of the model learns X and Y coordinates, a goal-independent representation of the environment, and this provides the flexibility necessary for DMP by allowing navigation to arbitrary goals. The complete model combines coordinates with the actor-critic architecture and accounts for the performance of rats in the RMW task, including the reversal, and in the DMP task.

The question posed at the end of chapter 2 was: how might place cell activity be useful for navigation, without containing all the spatial information necessary for navigation? The model of place cells considered in that section, introduced more as a model of what place cells don't do than of what they do, in fact provides an excellent representation for learning values, actions and coordinates. This concurs with previous work investigating the usefulness of the place cell representation (Dayan, 1991). Reinforcement learning methods such as the actor-critic are infamous for the large numbers of training trials required for learning, which in most applications run to the thousands. With place cells as an input representation, the actor-critic learns the RMW task in about 10 trials.

As well as considering a somewhat standard application of TD learning, the actor-critic,

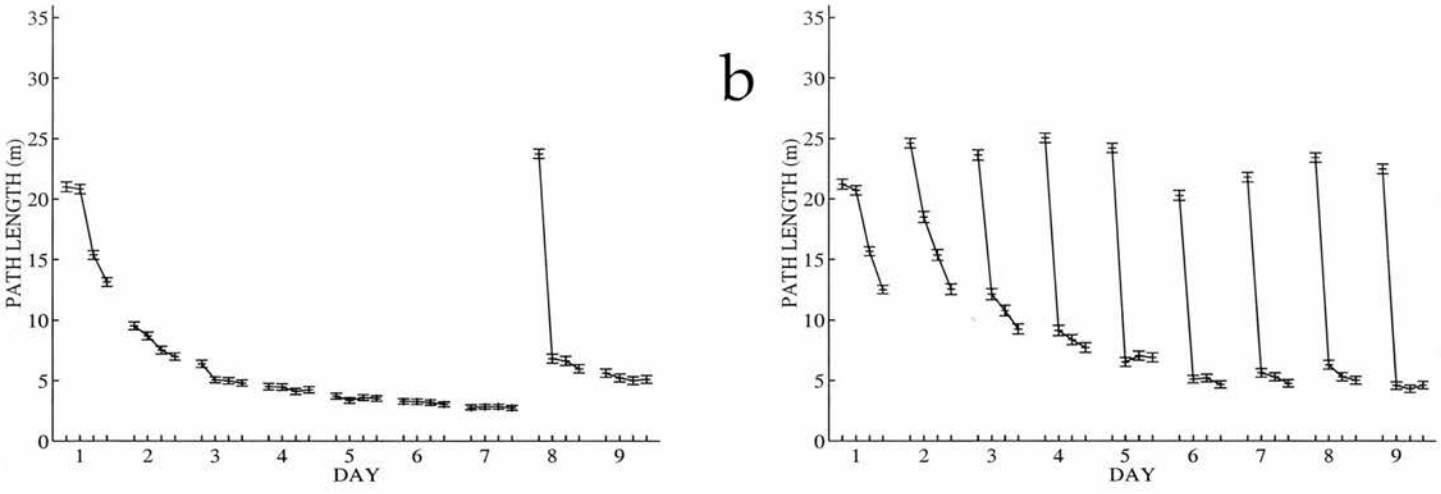


Figure 4.8: Performance of the combined coordinate and actor-critic model. For each data point, the mean and standard error in the mean are obtained from 1000 simulation runs. (a) RMW task, in which the platform occupies the same location. The combined model captures both acquisition, producing direct paths after around 10 trials, and reversal, producing rapid adaptation to the change in platform position on day 8 (see figure 4.1a). (b) DMP task, in which the platform remains in the same position within a day, but occupies a novel position on each new day. The combined model captures the acquisition of one-trial learning – the improvement within each day is gradual early in training, but becomes a one-trial improvement by day 6. The model provides a good match to the data (figure 4.1).

a novel application of TD learning has also been presented in the form of a network that learns consistent coordinates in an environment. This learning is independent of exploration strategy, and so is found to be extremely fast, with smooth coordinates acquired after about 16 trials, even though all the while control is being optimised to a different goal every four trials. Moreover, the coordinates learned are stable, despite being learned from relative information. The problem of global consistency is a general one that affects all navigating systems which use self-motion information to build map-like representations. The solution presented here partners a statistically efficient learning algorithm, TD learning, with the stable, allocentrically defined representation of the environment that hippocampal place cells provide.

Implications for the tasks

What does the model tell us about the spatial tasks themselves? First, since the actor-critic component can capture acquisition performance of rats in RMW, this acquisition does not provide evidence for a **Cognitive Map** (O'Keefe and Nadel, 1978; Morris, 1982). The actor-critic is not the first model to provide a non-mapping account of the task (Zipser, 1986; Wilkie and Palfrey, 1987; Burgess et al., 1994; Brown and Sharp, 1995; Blum and Abbott, 1996). It is, however, the first to incorporate a principled solution to the distal

reward problem, the critical component of which is the temporal difference (TD) learning rule. This solution is quite general, since nothing is assumed about the topology of the environment (beyond the structure implicit in the place cell representation), and so the actor-critic has the potential to learn in more complex environments, such as environments with barriers.

Second, the model demonstrates that it may be dangerous to conclude, as in a recent review of models of navigation by Trullier et al (1997), that *metric* navigation methods subsume *topological* navigation methods. The DMP task can be solved using metric information supplied by the learned coordinates, but the model knows very little about the topological structure of the environment, and this is its principal weakness. Likewise, other demonstrations of navigational ability – such as execution of paths in the dark (Collett et al., 1986) or the taking of short-cuts (Menzel, 1973; Gallistel, 1990) – provide evidence for the use of metric information, but not necessarily for the learning or use of topological information about environments. Few spatial tasks demand even coordinates, and a challenge for the future is to explore whether rats use still more sophisticated (*ie* topologically richer) representations of space.

Limitations of the approach

The work in this chapter is based on many simplifications, the most obvious of which is the simple, radially symmetric gaussian place field model of equation 2.1, but there were others. A way by which the coordinate controller might suggest a value to the critic was not included – and so the critic itself becomes inaccurate as one-trial learning is established. This may be justified on the grounds of parsimony: there is little evidence to constrain the choice of mechanism either for this, or for the closely related issue of learning 'set', *ie* the information about the task that a rat acquires as it finds the platform changing position each day. Furthermore, 'set' learning is clearly incomplete, since on the first trial of each new day, normal rats continue to revisit the position of the platform on the previous day, even though this always incorrect (Steele and Morris, 1999).

The problem of learning to navigate to goals in many different environments has been avoided by assuming a single place cell representation and a single environment. This makes it difficult to predict what would happen after, for example, transfer to a novel environment. In fact, the data does not clearly indicate how place cell activity would change under these conditions, and any prediction of the model would be based upon such data.

Other unmodeled aspects include the formation of place fields themselves, and the possibility that place fields change during either task. Also, the search strategy of the simulated rats was based on a random walk, commencing at the starting point and ending at the platform.

A better strategy would be to search all areas of the pool more uniformly, and this may indeed be what the experimental rats did (thus achieving somewhat better trial 1 performance in DMP than the model, according to figures 4.1b and 4.8). The key difference between the two strategies lies in not returning to previously searched areas. In fact, the actor-critic has the potential to learn such a strategy, if punishments are associated with moves which do not take the rat onto the platform. In this case, novel areas will appear more attractive than previously searched areas.

4.4.1 Relationship To Experimental Data

The Hippocampus

In the model, weight changes due to navigational learning occur downstream of hippocampal place cells. Consonant with this, Steele and Morris (1999) find that, after 9 days of pre-training on DMP, animals can, at short memory delays, continue to perform one-trial learning to novel platform positions during pharmacological blockade of NMDA receptors in the hippocampus. However, 9 days is long enough to learn a coordinate system, and so the experiment of Steele and Morris does not distinguish between models in which coordinate-like information is stored inside the hippocampus, and models in which it is stored outside.

An interesting issue raised by the model concerns the explicit memory for the current goal location, demanded by the coordinate model. Evidence exists suggesting that goal memory may be a dissociable computational factor in navigation. Steele and Morris (1999) find that, after 9 days of pre-training, animals in which hippocampal synaptic plasticity has been blocked by an NMDA antagonist show a delay-dependent impairment during DMP. That is, trial 2 performance in DMP is impaired if, and only if, the delay between trials 1 and 2 is long (20 min or 2 h; short delay was 15 s). Within the framework of the model, this delay corresponds to a selective disruption of the goal coordinate memory. Moreover, the data suggests that the normal operation of this goal memory is dependent on the normal operation of hippocampal NMDA receptors.

The Actor-Critic

As discussed in section 3.6, the actor-critic is a general learning scheme that has been used to model phenomena in classical and instrumental conditioning that are likely to be largely independent of the hippocampal formation, but which also may have an involvement in navigational learning. In line with a recent model (Montague *et al*, 1996), the dorsal striatum may play the role of the actor, while the ventral striatum plays the role of the critic.

Both the ventral and dorsal striatum of the rat receive outputs from the CA1 hippocampal subfield, an area where place cells are found (Wiener, 1996). However, little is currently known about the activity of these systems during navigation, or how or where the values may be stored.

The Coordinates

There is no evidence as yet for the neural implementation of the coordinate representation. However, the phenomenon of dead-reckoning is well documented in many animals (Gallistel, 1990), and strongly suggests both that a coordinate representation of some sort exists, and that neural mechanisms exist to perform simple vector subtraction. The particular X and Y coordinate representation used is extremely simple – and was chosen to demonstrate the problem of building globally consistent coordinates from relative self-motion information regardless of what sort of coordinate system is being used. It is not expected that a Cartesian representation is necessarily present in the brain.

4.4.2 Relationship To Other Models

The two key issues separating models of navigation are, from a neural perspective, the extent to which the hippocampus itself solves the navigation problem, and, from a computational perspective, the generality of the suggested control scheme. Both actor-critic and coordinate components use the hippocampus only for a representation of state (*ie* location). The actor-critic is a completely general control mechanism, working in environments with arbitrarily complicated shapes and reward contingencies, but is fairly inflexible. The coordinate model is flexible, but specialised to navigation in a restricted class of environments.

Blum and Abbott's (1996) model (see also Abbott and Blum, 1995; and Gerstner and Abbott, 1997) is closely related to dynamic programming. They propose that place cells express a decodable population code for position, and that subtle changes in the population code, due to the operation of temporally asymmetric Hebbian synaptic plasticity between place cells in field CA3 while the rat is swimming, can be interpreted as reporting at each location the average swimming direction that takes the rat to the goal. This essentially performs one step of the dynamic programming technique of policy improvement, starting from a random policy (Dayan and Singh, 1996). However, for general control problems, just one step of policy improvement is inadequate; even in the RMW task which they modelled, it was necessary to include a reinforcement process which modulated the Hebbian plasticity, in a manner similar to Brown and Sharp (1995).

Gerstner and Abbott (1997) extended Blum and Abbott's (1996) model to the case of nav-

igation to multiple goal locations. In their model, the (remembered) position of the goal modulates the activities of place cells, allowing the connections between the single set of place cells that are active in an environment to store the swimming direction appropriate to the multiple goals. Having learned synaptic weights appropriate for a few goals, navigation to novel goals is possible by interpolation. The model might use this feature to solve DMP, even in the face of the pharmacological blockade. However, there are various counts against the model. First, the modulation of place cell activity by goal position is not observed – in fact there is evidence against it (Speakman and O'Keefe, 1991). Second, both versions of this model embed the whole problem for navigation in the hippocampus proper, in the connections between CA3 cells. This is hard to reconcile with the results of Bannerman *et al* (1995) suggesting that plasticity in this region may not be necessary to learn a watermaze task in a novel environment, where place cell activity might be very different. Third, one of the key computational operations in the models is population decoding of the position of the rat that is encoded in the activities of the place cells. Calculating this requires knowledge of something equivalent to coordinates in the environment, that is, *a priori* knowledge of the location (in some coordinate system) of the centre of each place field. Some additional, unspecified scheme for learning these coordinates consistently across the environment is essential.

Like the actor-critic system, Burgess *et al* (1994) and Brown and Sharp (1995) have also suggested schemes in which place cells play the more limited role of providing a reliable code for space. Both papers consider an RMW-like task which presents a distal reward problem. Burgess *et al* (1994) use the output of place cells to construct subicular cells with extended place fields, which in turn are used to learn postulated goal cells, which fire across the extent of an entire environment, performing a job like the actor. Learning of the goal cells only happens when the animal actually reaches the goal, but this is sufficient because the extended range of the goal cells means, in effect, there is no longer a distal reward problem. If by some means the firing of goal cells for different goals could be distinguished, it is possible the model could also address the DMP task, by having a subicular cell for every possible goal. However, the use of large firing field representations in this manner raises a number of issues. First, if the subicular cells that fire when the animal is at the goal do not cover the whole environment, there will be places for which the animal will not learn appropriate actions. Second, the mechanism which generates large subicular fields can be expected to learn more slowly than TD-based value learning, and to impose greater constraints on exploration than TD-based coordinate learning, since it attempts to produce a smooth, monotonic function of distance in the subicular cells by essentially averaging over place cell activity traces for each subicular cell (*ie* for each potential goal). Third, the model does not use a general learning scheme for control, and so can only accomplish tasks such as avoiding obstacles by making detours that are significantly larger than necessary and

which, for inconveniently located barriers, may not work at all. Brown and Sharp (1995) present a simpler model in which place cells are associated with responses, and in which learning is gated by reward. However, as noted in the introduction, the model relies on a trace-like learning rule which is likely to be a very inefficient way of learning predictions compared to the TD learning rule used in the actor-critic model. The model does, however, suffer the same limitations as the actor-critic with respect to the learning of a DMP task.

The problems involved in learning a coordinate system have been addressed by Wan *et al* (1994). In their model, coordinates are represented by an extra-hippocampal path integration module that operates more conventionally, representing coordinates with respect to some current point of origin. Their model demonstrates how place cell firing might come through learning to be independent of sensory information, at least for a short while, relying instead on input from the path integrator. It also addresses the inverse problem of what happens when the path integrator becomes invalid, as for example on each new trial of a watermaze task, because the path integrator learns to set itself by the output of place cells. In a completely novel region, a new origin is selected and new coordinates laid down. However, if previous experience is of value to the animal, it must return to areas of the environment where place cells can correctly set the path integrator; so for example trial 2 of a watermaze task could not produce any learning until a familiar area was traversed, so throwing away potentially valuable experience, as well as constraining the animal's search. The TD-based model avoids both shortcomings by directly tackling the problem of inconsistent coordinates.

Finally, a quite different view of hippocampal function from that taken by the models discussed so far is that the hippocampus is directly involved in some forms of flexible processing, for instance manipulating sequences of mnemonic or spatial information (Zipser, 1986; Levy, 1996) or performing complicated computations, as in the demonstration of transitive inference (Bunsey and Eichenbaum, 1996). Although direct experimental support for this view is lacking, it is not possible, on the basis of current evidence, to rule it out. However, transitive inference may be a case in point, because working out a global order from local relationships is a similar task to that of calculating globally consistent coordinates from local dead-reckoning information. It is possible that the hippocampus computes the inference directly; it is also possible that downstream systems make the computation, but rely on the hippocampal representation to do so. With regard to navigation tasks, it has been demonstrated here that although the observed activity of place cells appears limited, it makes sense if used in the right system with the right learning rule. Indeed, by this model's account, the very characteristics that make place cell activity seem so redundant – namely localisation, directional independence and stability – contribute most to their suitability within a navigational learning context.

4.4.3 Predictions Of The Model

On the basis of the model, the following three predictions can be made.

1. **Placement trials should support DMP, once rats have acquired one-trial learning**. After a certain amount of training, rats should have a system that specifies the coordinates of any location they occupy. This implies that, by this stage of learning, mere placement on a platform in a novel position might be sufficient to allow asymptotic performance of the next trial. This prediction was in fact tested in a behavioural experiment described in the following chapter.

2. **Rats for which hippocampal synaptic plasticity is blocked, but only after place fields have been established in an environment, should be unimpaired in learning a RMW task.** The model suggests that the actor-critic is located outside of the hippocampal formation, and just uses information from the place cells as a representation of state. Therefore, provided the place cells have been established (*eg* during a latent learning period of some sort), actor-critic learning should progress normally. The complication is learning set behaviour - if blocking plasticity prevented the animals from learning the nature of the task, this too would have to be ensured during a pre-training period.

3. **Rats for which hippocampal synaptic plasticity is blocked, but only after place fields have been established in an environment, might also be unimpaired in learning a DMP task.** If this was found to be true, it would suggest that the coordinate system (in particular, cells X and Y in the model) is located outwith the hippocampus. An impairment, on the other hand, would suggest that coordinates are located within the hippocampus. The experiment of Steele and Morris (1999) did not distinguish between the two alternatives because synaptic plasticity was blocked only after extensive pretraining (which provided the one-trial learning data which we have modeled). The same considerations apply for this prediction as for the previous one, in terms of establishing place fields, and acquiring the learning set.

Chapter 5

Experiments Further Investigating One-Trial Learning

5.1 Introduction

It was argued in chapter 4 that a minimal model of hippocampal dependent navigational learning captures rats' performance in two watermaze tasks: a reference memory task, and a computationally more challenging delayed match-to-place task. The question arises whether the model is a comprehensive model of rats' navigational abilities, or whether an appropriate experiment would reveal navigational learning of which the model is incapable.

One prediction of the model is that once coordinates have been learned, one trial learning should be evident after mere placement of the rat on the platform, a qualification being that the rat is able to understand the significance of the platform during placement trials. The principal weakness of the model, as noted in chapter 4, is that a coordinate-based controller of the sort utilised in the model cannot specify actions in an environment with a complex topological structure, such as an environment with barriers. It therefore predicts that rats should be unable to accommodate barriers in one-trial learning navigation tasks. This chapter examines both issues in detail. First, a review is presented of the experimental literature concerning navigation, focusing on experiments with rodents in mazes, with the capabilities and limitations of the model in mind. Second, two experiments are presented which address directly the questions of whether rats exhibit one trial learning after placement, and in the presence of barriers.

5.2 Literature Review: behavioural studies of navigation

The early history of psychology is rich in experiments with rats in mazes. Later periods have witnessed a tendency to simplify experimental designs, in tandem with the introduc-

tion of sophisticated techniques for measuring and manipulating neural mechanisms, from lesions to single-unit recording. In general, the factors influencing behaviour are controlled more rigorously than in the early experiments. Although early experimenters were clearly aware of the possibility of multiple explanations for the behaviour of animals in their experiments, they did not develop the kind of control over cues and over motivation evident in, for example, the modern watermaze task. Nevertheless, many of the early studies stem from an interest in spatial abilities more closely aligned to a computational interest than the modern studies.

In the interests of economy, this review will not consider experiments examining how an animal's understanding or sense of *where* things are depends on environmental cues. Thus, for example, we will not consider the interesting experiments of Collett *et al* (1986), in which the search strategy of gerbils attempting to find a previously visited but hidden food source was manipulated by moving or removing landmarks in the environment. This exclusion is justified because the computational questions we are interested in concern how animals learn to choose appropriate actions in an environment with which they are familiar. By contrast, changes in the position of landmarks constitute changes to the environment, and may reveal more about how an animal recognises and represents places than how it learns to choose appropriate actions at different places. There is, in fact, evidence that these two aspects of spatial learning are separable, and this was discussed extensively with regard to the hippocampal representation of space, in chapter 2. In particular, the results of, for example, Collett *et al* (1986) may be better understood in terms of how place cells come to fire where they fire, than in terms of how place cell firing leads to navigational behaviour. Although many of the experiments reviewed here were conducted at a time before place cells were discovered, it is this kind of distinction to which I shall attempt to adhere.

5.2.1 The Early Days

Some of the original sources of the work described in this section either were not available, or else were published in a different language to English (German). Therefore, while the original references are given in the text, a portion of the description of some of the work follows reviews by Maier and Schneirla (1935) and Tolman (1948; 1949; 1951).

Goal gradient

Hull (1932) introduced the "goal gradient" hypothesis, whereby learning in a maze was supposed to proceed more rapidly closer to a goal. However, the order of elimination of maze errors had been a contentious issue before Hull's publication. The issue is potentially

of interest in the light of the reward-based component of the model of the previous chapter. In that model, correct estimates of the value of locations tend to appear close to the goal first, and spread out to the rest of the environment only later. The effect on the paths generated by the simulation was not as clear, mainly because of the contribution of momentum in the watermaze to the simulated animal's trajectory. However, many early mazes used paths clearly broken up by distinct choice points, thus perhaps reducing the effect of momentum. In such cases, an explanation of learning in terms of a value function would predict that correct actions would appear first at choice points closer to the reward site, and only later at choice points further away, *ie* the actor-critic would predict a backward elimination of errors.

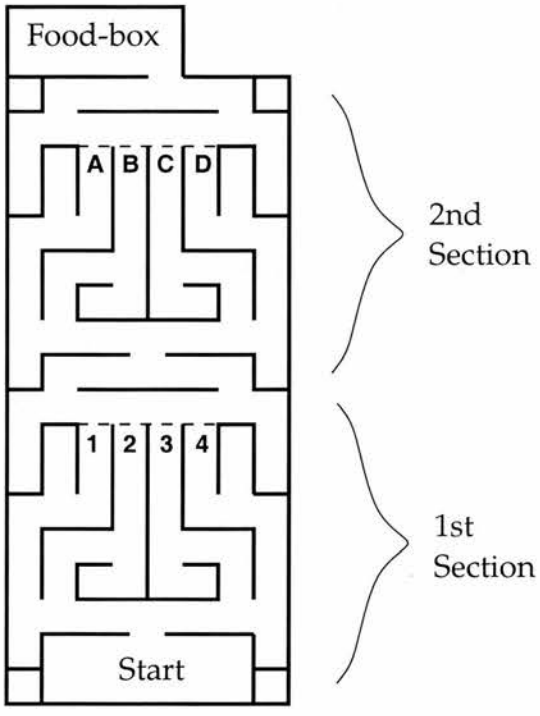


Figure 5.1: The maze of Borovski (1927). The two sections of the maze are similar in structure. The route through each may be altered by removing or inserting barrier sections into the alleys 1, 2, 3, 4, and A, B, C, D. After Maier (1935), p. 401.

Such a prediction appears to be supported by the evidence. Borovski (1927) conducted a maze learning experiment with rats using two identical, complex mazes, one of which (section 1) led to the other (section 2), which in turn led to the goal (figure 5.1). In the experiment, the direction of the goal (food) would not tell an animal which paths were dead-ends and which were not. This is important, because we do not know that directional information was not available to the animals, *eg* in the form of olfactory cues, so as to render the selection of paths heading towards the goal trivial. Three groups of animals were trained on the maze task. The first group (control) were repeatedly given trials without any changes made to the mazes. The second group were trained with repeated changes made,

from trial to trial, to section 1, such that different paths led to dead-ends on each trial. The third group suffered similar changes, but to section 2, which was the closer to the goal location. Maier summarises the results as follows: first, the control group learned section 2 before section 1; second, the introduction of trial by trial changes, for this control group, after they had learned the task, was more disrupting if those changes occurred in section 2 than in section 1; third, the third group (changes made to section 2) made many more errors than the second group (changes made to section 1). The implication is that the learning in the first section (far from the goal) may be dependent on learning in the second (close to the goal), but perhaps not vice-versa.

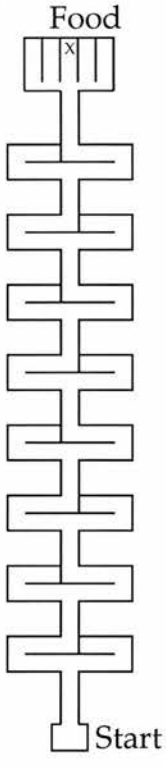


Figure 5.2: The linear maze of Buel (1934), in which directional information is irrelevant to navigational choices of action. After Maier (1935), p. 395.

Further support comes from an experiment which was careful to nullify the contribution of directional information (Buel, 1934). The apparatus presented an animal with a series of choice points arranged along a line, at each of which the goal was located directly ahead but the animal had to choose instead to go left or right (figure 5.2). At each choice point, one or other choice led to a dead end, which was out of sight at the choice point itself. The results revealed a strong influence on the action chosen at every choice point of the correct action at the very last choice point, *ie* rats tended to make the choice everywhere that was correct at the choice point leading directly to the goal. However, errors were eliminated backwards from the goal position. A drawback of the experiment is that, as shown in figure 5.2, the correct choice simply alternated between successive boxes.

Are these tasks likely to have been hippocampally dependent? It is possible that hippocampally lesioned rats could learn eventually to perform a sequence of responses, such as is required by both experiments. Thus, it is difficult to draw strong conclusions about the nature of navigational learning from these experiments, and indeed it might be easier to accept that simpler systems were doing the job (*eg* reward-based systems subserving extra-hippocampal procedural learning). Certainly this is the more attractive interpretation for those who set a premium on keeping theories of hippocampal function separate and distinct from theories about the rest of the brain. However, adherents to this interpretation have to accept that, in this task, the hippocampus was unable to make a contribution – and so presumably would predict that this task would be relatively insensitive to hippocampal damage. By contrast, theories such as the actor-critic model, which allow the hippocampal system to *work with* extra-hippocampal reward-based systems, expect to see characteristics of these systems, such as the backward order of error elimination, as well as characteristics of the hippocampal system, such as a facilitation of learning in these tasks due to the appropriateness of the hippocampal representation of space. Thus, normal animals would be expected to learn more rapidly than hippocampally lesioned animals. The relevant data is, sadly, unavailable.

One-trial learning

Perhaps the class of experiments most relevant to the questions heading this chapter are those that explicitly addressed one-trial learning in a navigational context. Unfortunately, there are clear problems with the interpretation of the results in each case.

Maier (1929) allowed rats to explore freely a maze consisting of a single path, but a portion of which offered three parallel tracks between two points, *ie* the path split three ways, and then all three met again and a single path continued. They were then taught to traverse paths from start to finish (where a food goal was located) that could include either of two of the parallel sections, but not the third. In the third phase, the food was moved to a position half-way along the prohibited section, and rats were simply placed there to experience the food. In the fourth and final phase Maier reports that 4 rats, starting from the same starting position as before, chose, at the appropriate choice-point, the previously prohibited path, and so found the food. The 3 remaining rats in this small study proceeded to the old food location, but then went back along the previously prohibited path, and found the food. The critical problem with the study would seem to be that local cues are available for the new goal position, and at both choice points where rats were observed to choose the correct path (*ie* on the way out from the start, and on the way back from the old goal), these cues were visible. Having seen the cues, the animal needed only to head towards them to exhibit the behaviour which Maier interprets as reasoning, and which we might have hoped to interpret

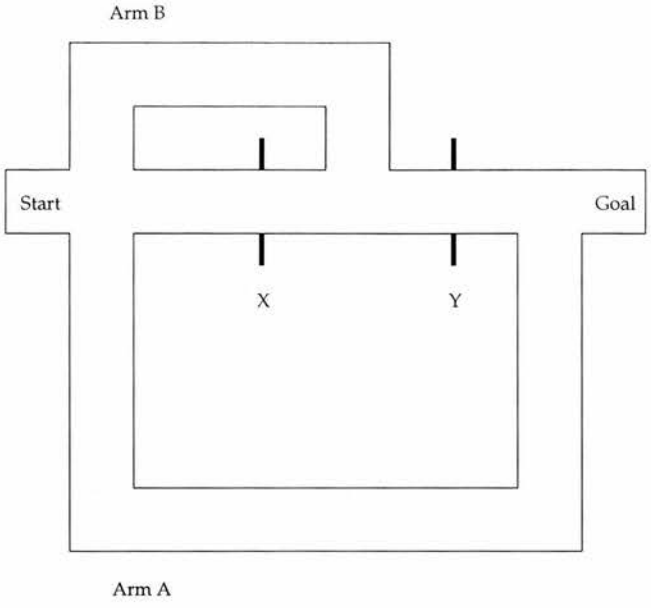


Figure 5.3: A schematic diagram of the experiment conducted by Tolman and Honzik (1932a). Rats could take one of three routes from a start box to a goal box. However, by placing a barrier at either X or Y, the central route could be blocked off, forcing the rats to choose arm A or arm B. Depending on where the barrier was, one of the arms was the optimal choice. Amended from Tolman (1949)

as one-trial learning. The further, trivial possibility also exists of directional cues from the food itself governing the animals' choice of action. Nevertheless, one conclusion to draw from this discussion is that a convincing demonstration of one-trial learning would need to not only avoid the problem of cues naturally associated with the goal (such as the smell of the food), but also avoid the problem of local cues whose association with the goal has been *learned*.

Tolman and Honzik (1930a) presented a one-trial learning result which is often quoted, but which is somewhat difficult to interpret. Rats familiar with a maze consisting of three paths to a goal were forced to choose a path to the goal after the most direct one was blocked (figure 5.3), one of which was either the shorter of two viable paths or the only viable path, depending on the position of the barrier. It is often reported that rats were able to choose the correct path *immediately after experiencing the barrier*. If true, this would indeed have constituted evidence for one-trial learning abilities of a fairly sophisticated kind, and beyond the capabilities of the model presented in chapter 4. In fact, the first attempt with this task was a failure. Only when the paths in the maze were elevated to allow the animals to perceive the maze was such a result achieved. Under these conditions, the rats were first allowed to explore the maze, and then were allowed to react to a barrier favouring the short indirect path. Only then were rats tested with the barrier in the position requiring the

longer path, at which point the majority of rats chose the third, longest path. Interpretation of the result is made difficult, therefore, by the apparently important role played by direct perception in the experiment.

Latent learning

One of the best known results of the period was the demonstration of "latent learning" (Blodgett, 1929; Elliott, 1929; Tolman and Honzik, 1930b), which was discussed in chapter 1. In particular, it was demonstrated that the mazes used for these studies trivialised the predictive learning problem, such that little can be infered from the studies about the extent of the rats' learning about the navigational environment (figure 1.1). As a more concrete example of the weakness of the experiments, it may be considered how the model of chapter 4 would fare on these tasks.

In fact, only the reward-based actor-critic component of the model need be considered (*ie* without coordinates), this component being sufficient to account, with just a small adaptation, for the latent learning results described. The mode of action selection in that model allows for a variable parameter (β in equation 3.17) determining to what extent the learned preferences of the model need influence actions. Using this parameter, it is possible for exploratory behaviour to mask learning. If an encounter with a dead end is taken to be a somewhat aversive experience, it would be possible for the model to learn preferences for appropriate action choices, while consistently choosing exploratory (*ie* "incorrect") actions. Having found food, exploration may be less attractive. Although the model does not incorporate a mechanism by which exploration is deemed more or less attractive, nevertheless the computational effects of a more focused behaviour can be simulated by an increase in the parameter β. In this way, the demands of changing rapidly from a non-goal-specific strategy to a goal-specific strategy are less stringent than the demands of changing from one goal to another. Note, however, that in general actor-critic learning, improvements in the policy would necessarily complement improves in predictions (*eg* the "bootstrapping" of value and policy learning described in section 3.4.5). It is the simplicity of the prediction problem in these "latent learning" experiments, and in particular the absence of a temporal credit assignment problem, that allows correct actions to be learned during an exploratory policy.

5.2.2 The Modern Era

By comparison with the tasks that have been discussed so far, modern spatial learning tasks tend to be both simpler in computational terms, but also more carefully designed in terms

of control over both the behaviour of subjects, and over the sources of information available to subjects during the performance of the tasks. Two classic examples of a simplified but controlled testing apparatus, the watermaze and radial maze, have been discussed at length already (sections 1.5, 1.6.1, 2.2.2, 2.2.3 and 2.2.4).

Placement-Learning Studies

Keith and McVety (1988) investigated the effects of platform placement in a kind of one-trial learning task in the watermaze. The specific form for this one-trial learning involved pre-training in a watermaze in one environment, before a single swimming trial in a watermaze in a novel environment. Unfortunately, and in contrast to DMP within the same environment, this is, if generalisation at all, then generalisation across environments and so theoretically problematic. This point is returned to below.

49 rats were pre-trained on an RMW task in the first watermaze, for 3 days (10 trials/day). For one lot of rats each single trial consisted of first placement (30s) on the platform, and then subsequently a conventional swim trial. However, four other groups of rats underwent different kinds of trials: swim-only trials, placement-only trials, unrewarded swim for a time yoked to that of one of the place-and-swim groups, or simple handling, *ie* no experience in the watermaze at all. Keith and McVety reported that placement-and-swim rats performed better than swim-only rats at this stage.

On the fourth day, the rats were given one trial in the second, novel watermaze. Those rats given place-and-swim trials first time round were now divided into three groups each receiving a different kind of trial: a place-and-swim trial, a swim-only trial or a place-and-swim trial but with the platform moved to its opposite location for the swim. All the other pre-training groups received place-and-swim trials. Note that only two out of all the groups of rats experienced the platform during pre-training *and* received an appropriately positioned placement trial before being required to swim the training trial. Keith and McVety report that these two groups showed significantly better latencies on this swim trial than all other groups. 13 out of the 16 rats in these two groups visited the correct quadrant first on this swim trial.

The prediction of our model in this task is difficult to make, for two reasons: first, the RMW task itself may or may not force coordinate learning (although in chapter 4 post-RMW reversal performance was modeled by coordinate learning). In any case such coordinate learning has only had 3 days in which to be learned (although 30 trials). The second, more fundamental difficulty is that any coordinate learning that does occur must then transfer to the new watermaze, and the prediction made by our model depends upon how hippocampal place cells generalise in this situation. Nevertheless, *if* coordinates can

be learned, and *if* place cells generalise in the required sense between watermazes (and if the very act of placement is not too confusing for the animals), then our model does predict one-trial learning through placement. The results do in fact indicate some learning but without comparison to a swim-swim group, or further training, it is hard to claim that this is asymptotically good performance acquired in one trial. This and the similar point raised by Chew *et al* (1989), that the swim paths do not appear to be direct, are important because of the trivial possibility that *some* improvement in performance might arguably be mediated by non-spatial means, *eg* by rats heading vaguely in the direction of a distal cue that appeared to be near the platform and so increasing the probability of a short latency.

Using a task design similar to DMP, Whishaw (1991) presented evidence for generalisation across successive watermaze tasks involving the same environment but different platform positions, reporting that eventually rats display one-trial learning of novel platform positions. This result followed previous work which used a set of four possible platform positions but hence not novel platforms after day 4 (Whishaw, 1985a). Whishaw (1991) also examined the possibility that mere placement of rats on a platform would support one-trial learning. This is very close to the prediction claimed for the model of the previous chapter.

An initial experiment established various parameters, including the fact that a single placement was as effective as five placements in a row in facilitating a subsequent swim. It also revealed a significant effect of day over the 10-day training period, in line with our model's prediction that only after coordinates are learned would placement prove useful. The second, main experiment was a rather large study (230 rats). The basic training each day consisted of a 30s placement on the platform, followed by two conventional swim trials (though the second always used the same starting position as the first). However, several different groups incorporated: various delays between placement and the first swim (0s, 5s, 5min, 1h, 4h, 24h), placement in an incorrect location (*ie* with respect to the two swim trials following), or no placement at all. This last group, as well as providing a control, allowed examination of the effect of the first swim trial on the second. Training lasted for 16 days, of which the first 15 days were in a watermaze in one environment, but the 16th day used a different watermaze in a novel environment.

The results for the first 15 days appear to corroborate those of Keith and McVety (1988) and because of the number of rats used, many effects were found to be significant. The mean latency on the first swim trial was significantly better for all placement groups (except the 24h delay one) when compared to wrongly placed or not placed groups. This performance was, however, significantly worse than the trial 2 performance of the non-placed group, implying that placement was not as effective as swimming in supporting one-trial learning, inviting similar criticisms as the Keith and McVety (1988) study. The best, placed mean latency (for 1.46m diameter pool) was 21.9s; the mean for non-placed rats was 40.6s.

There are reasons to be suspicious, however. First, the statistical comparisons were made on the basis of all the data, *ie* without excluding the period of pre-training during which, according to our model, coordinates would be learned. A figure in the paper shows that placement and swim latencies come closer together towards the end of the 15 day period. The second reason stems from an analysis of the probability of returning to the previous day's platform position by placed rats. This is an interesting question because swimming rats in a DMP task revisit the previous day's platform position on their first-trial swim to a novel platform (Steele and Morris, 1999) and while placed rats experience the new location of the platform, they do not experience the absence of a platform at the previous day's location. Indeed, Steele and Morris (1999) report that even after 17 days of novel platforms rats still return to the previous platform location – suggesting it might be hard for placed rats to learn to do otherwise. Whishaw reports that the probability of swimming first to the previous platform position increases smoothly with the delay between placement and swim, but even for the 0s delay group this probability is more than half. What is missing, however, is an estimate of the same probability for trial 2 for the swim-only rats, which if it were much lower would offer an alternative explanation for why placement did not appear to be as effective for one-trial learning as swimming.

An astonishing further result was obtained on day 16, which took place in a different water-maze in a novel environment. Placement resulted in significantly better performance compared to non-placed or wrongly-placed rats (best placed mean: 16.2; non-placed: 34.6). It must be concluded from this result that there clearly is transfer between different water-mazes. However, it is difficult to predict the corresponding performance of the model of chapter 4, since as discussed in that chapter not enough is known about the activity of place cells in the new environment.

Navigation in the presence of barriers

An experiment with cats (Poucet, 1983) has addressed the question of which of direction and traversable distance are most important in determining an animal's choice of detour around a barrier to get to a goal. Animals were tested indoors in a room rich in distal cues, and in four conditions, in each of which the animal had to take a detour around a barrier to obtain a food reward, and essentially was made to choose to go by one of two routes around the barrier. However, the barriers and goal position were different in each of four conditions, with specific consequences for the detour routes (figure 5.4). In the first, there was a clearly shorter path which also involved the smaller divergence away from the direction of the goal. In the second, the lengths of the paths were equal, but one path diverged less than the other. In the third, the divergences were equal, but one path was shorter then the other. In the fourth and final condition, the path which was shorter

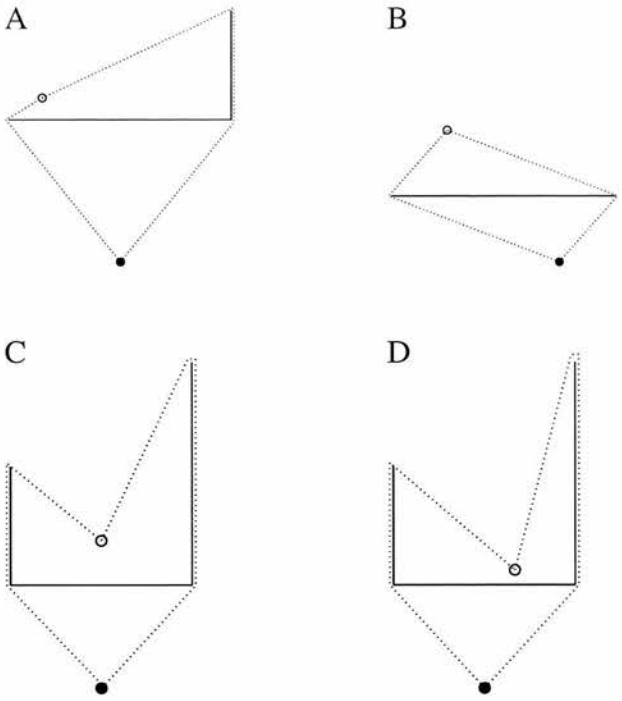


Figure 5.4: The four navigation conditions in Poucet *et al* (1983), in which the solid circle is the start position, open circle is the goal, solid line is a barrier, and dashed lines are possible paths: (A) distance and direction to goal are in agreement about which path to take; (B) distances are equal but direction differs, *ie* the left path involves least divergence from the bearing of the goal; (C) directions are equal but distances differ; (D) direction and distance information are in conflict, because the most divergent path is actually the shortest. After Poucet *et al* (1983).

incurred a larger divergence away from the direction of the goal, and vice versa. In this condition, therefore, directional information, such as might be afforded by a coordinate system, and distance information, such as might be afforded by an actor-critic's estimate of value, are brought into conflict. A further factor in the experiment was that each condition could occur with either a transparent barrier, allowing the goal to be perceived, or with an opaque barrier hiding the goal.

The results were obtained by averaging over the performance of the animals during 10 test trials of each condition, with both visible and hidden goals. Therefore, the question of one-trial learning was not addressed. Before the 10 test trials, animals were given two rewarded trials, in which they were forced to take each of the two possible routes. The visible goal produced a tendency to favour the least divergent route, except when this route was the longer. More interestingly perhaps, the hidden goal case produced significant tendencies to favour the less divergent and shorter route when these were the same, the less divergent route when the lengths were equal, and the shorter when the divergences were equal. In the fourth condition, in which the two types of information are brought into conflict, it is reported that the shorter route was taken. However, here the authors additionally provide some data of interest to the question of one-trial learning – namely that on the very first trial

of this fourth condition, five out of eight animals chose the less divergent but longer route, before coming to choose the shorter route in later trials. If any conclusion can be drawn from this it is that while there was a tendency for animals to learn to take the shorter of two routes to the goal, this learning was apparently not one-trial learning, but of the gradual nature one might expect from the actor-critic model. However, interpretation is difficult not only because of the lack of data, but also because: (1) it is not known to what extent the change in the shape of the barrier between conditions affected the animals' tendency to generalise across conditions, *ie* exhibit one-trial learning; and (2) because of the constant starting position, a rather trivial strategy of estimating the duration of a route could (at least in principle) support correct navigation in the task. A similar experiment was conducted with dogs, with similar results, however the interpretation problem is even greater for this experiment, since each condition was tested in an entirely different place (Chapuis, 1983).

5.2.3 Conclusion

A number of experiments has investigated questions closely related to those of interest to this thesis. Modern experiments strongly suggest that placement learning may be possible, although the question of whether one-trial learning can be supported by placement has not quite been settled, with the lingering possibility that placement may improve search strategy but never as well as a full swim trial. As for barrier experiments, most of the early experiments tackled interesting questions but, either through a lack of control over experimental variables, or through mixed results, fail to provide firm answers. Hardly any research of a similar sort has been conducted in the modern era, despite the availability now of appropriate testing protocols.

Therefore, the following two experiments appear well motivated. The first addresses the issue of whether, after a period of DMP pretraining, placement can support one-trial learning. The second addresses the issue of whether one-trial learning is possible in a familiar environment containing a barrier, *ie* in an environment in which a pair of coordinates by themselves cannot specify the correct actions to take.

5.3 Experiment 1: One-Trial Learning After Platform Placement

5.3.1 Aims and Methods

The aim of this experiment was to test the prediction made by the model of the previous chapter, that after a sufficiently long period of pre-training on a DMP task, mere placement

of a rat on a platform in a novel platform position should support one-trial learning, *ie* direct paths on the second trial.

Subjects

The subjects (N=12) were experimentally naive male Lister hooded rats.

Apparatus

The basic apparatus was an open field watermaze (Morris, 1984), consisting of a large circular tank (diameter 2.0m, depth 0.6m) containing water at 25 ± 1^{o} C. The water was made opaque by the addition of liquid latex (Cementone "Cempolatex") which prevents a swimming rat from seeing the escape platform (see below), and facilitates tracking of swim paths. The pool was located in the centre of a room containing various prominent extramaze cues (wall posters, wall cupboards, a bunched set of white curtains and a large metal frame). The room was diffusely illuminated by four floodlights located in the corners of the room.

Swim paths were monitored by a video camera mounted in the ceiling, the signal from which was relayed to a video recorder and from there to an image analyser. The x and y coordinates of the rats' position were sampled at 10 Hz by an Archimedes computer running the "Watermaze" application (written by R. Spooner).

A solid white cylinder of diameter 11cm., the surface of which rested approximately 2cm. below the water level, formed an escape platform which could be placed anywhere inside the pool.

Procedure

Pretraining

Rats were handled for several days before undergoing pretraining. Each rat then received 4 trials per day, for six days (figure 5.5). The platform was placed at a novel location on each day, and remained in that location for all four trials of the day.

For days 1, 2, 3 and 5, the daily procedure was as follows. At the beginning of each trial, the rat was taken from its temporary cage at the side of the room, and placed in at the edge of the pool facing the wall, at one of four cardinal starting positions (north, east, south or west) chosen pseudo-randomly. A clock was started manually once the rat was in the pool. If after 120s, the rat had failed to move onto the platform and remain there, the clock was

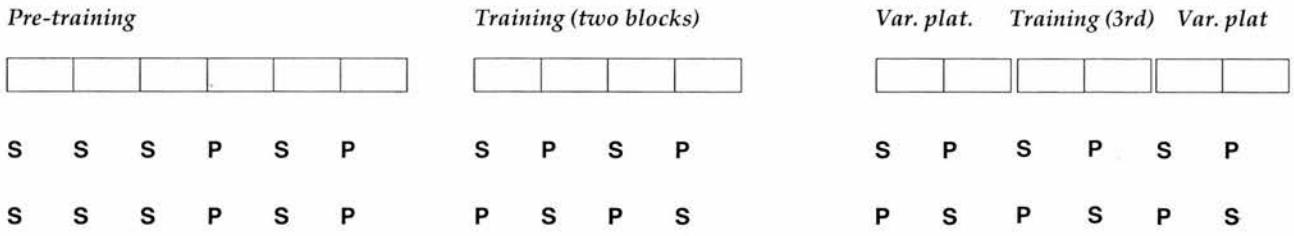


Figure 5.5: Experiment 1 protocol: While during pretraining both of two groups of normal rats received identical trials, during the training and variable platform phases, one group received a swim trial 1 (denoted 'S') while the other received a placement trial 1 (denoted 'P'). Trials 2, 3 and 4 were always swim trials for all rats. On each day, the platform occupied a novel position (and on variable platform trials it was moved to a second novel position for trial 2).

stopped, a latency of 120s was recorded, and the trial was ended. In this case, and where possible, the rat was guided (using the experimenter's hand) toward the platform, where it climbed on and remained, otherwise the rat was picked up and placed on the platform. Each rat remained on the platform for 30s. It was then transferred back to the side of the room (either into its cage or onto a towel lying next to its cage). The next trial of the day for the rat began a few seconds later. Rats were run in this manner one at a time, through all the trials of each day.

For days 4 and 6, the procedure differed in that the first trial of the day was a placement trial rather than a swim. For placement, each rat was taken from the side of the room and first held for about a second in a bucket of water at pool temperature, before being placed onto the platform for 30s (the same amount of time as at the end of swim trials). The three remaining trials of the day were normal swim trials.

Training

There were 6 days of training: days 7-10 and 13,14. Similar procedures to those used in pretraining were continued during the training phase. Rats received alternating swim days (S; all four trials being swim trials) and placement days (P; the first trial being a placement and the three remaining being swims). However, rats were split into two groups: one group began with S, the other with P (figure 5.5). Strict alternation was pursued because there was no perceived merit in attempting to keep the *type* of trial to be received a secret from the rat, if indeed a rat was able to learn to predict such a thing.

A further complexity was the sequence of platform positions used. Previous studies in the laboratory had indicated the importance of adequate counter-balancing of platform positions, particularly since some platform positions were particularly easy for rats to find quickly "by accident" (*eg* on trial 1 of the day). Following previously used procedures (Steele and Morris, 1999), rats were split into six pairs, each containing one S and one P rat. Each pair underwent a different sequence of platform positions in a counterbalanced

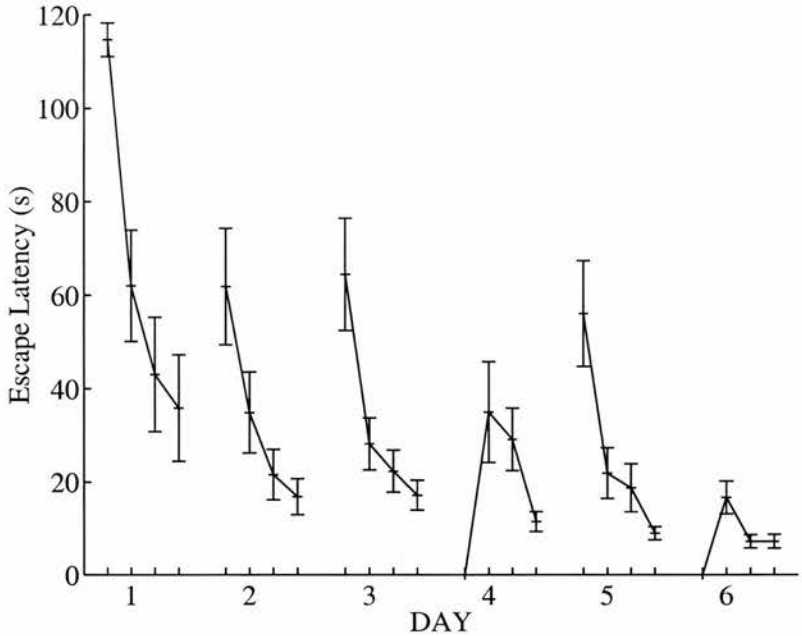


Figure 5.6: Pretraining: mean latencies (N=12) for the 6 pretraining days. Placement trials are indicated with a latency value of zero.

manner. The result of this organisation was that: (1) on any given day there was an appropriate range of platform positions; (2) every platform position was experienced by an S and a P rat on the same day; (3) over the course of training each rat experienced a balanced set of S and P days; and (4) the total number of "easy" and "hard" platform positions was balanced between the S and P groups.

Variable platform control

On days 11,12,15 and 16, rats received a variable platform control test. Maintaining S and P alternations, rats were given a trial 1 as during training. However, the remaining 3 swim trials were then conducted with the platform having been moved to a novel location. This is an important control for the possibility that enhanced performance on trial 2 is simply an effect of it being the second trial of the day (*eg* as if the first trial served merely to rouse the rats into action), or in fact any other non-spatially-specific feature. The sequences of platforms in this phase were balanced in a similar manner to those of training.

5.3.2 Results

Pretraining

Rats gradually acquired one-trial learning, as evident from figure 5.6. Moreover, on days 4 and 6, placement apparently did not appear to disrupt performance to any great extent.

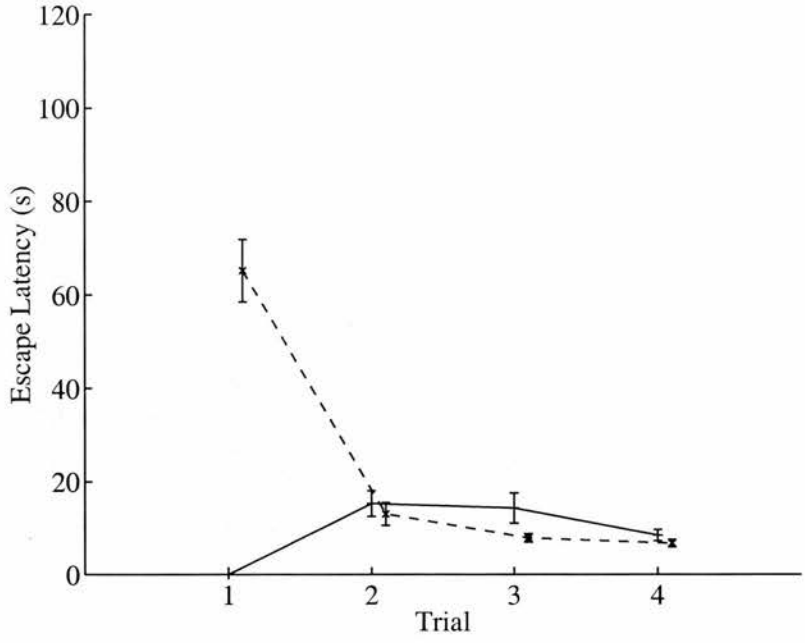


Figure 5.7: Training: mean latencies (N=12) over 3 blocks of training, for swim days (dashed line) and placement days (solid line). Note that the platform occupies a novel position at the beginning of each day.

Training

During training sessions, the pattern of latencies within days was similar to the pattern identified in previous chapters as one-trial learning (figure 5.8). Mean latency on trial 1 was 65.2 ± 6.7 seconds whereas on trial 2 it was 15.3 ± 2.8 seconds for placement days, and 13.0 ± 2.4 seconds for swim days. A repeated measures analysis of variance (ANOVA) with swim/placement as a within subjects factor revealed no significant difference for trial 2 performance [$F(1,11) < 1$] or indeed for trials 2, 3 and 4 together [$F(1,11)=3.45$; $p > .05$]. It was confirmed, however, that on swim days performance on trial 2 was significantly better than on trial 1 [$F(1,11)=75.5$; $p < .001$].

Variable Platform Control

Moving the platform after the first trial but before the second trial disrupted rats' trial 2 performance. A repeated measures ANOVA with trial 1/trial 2 as a within subjects factor revealed for swim days no significant difference between the two trials [$F(1,11)=1.15$; $p > .3$]. As during training, there was no significant difference between swim and placement days for either trial 2 performance only [$F(1,11) < 1$] or for trials 2, 3 and 4 together [$F(1,11) < 1$]. Overall, the variable platform was a highly significant factor, as revealed by repeated measures ANOVA over training and variable platform control data combined, with swim/placement and stable/variable entered as within subject factors [$F(1,11)=73.5$; $p <$

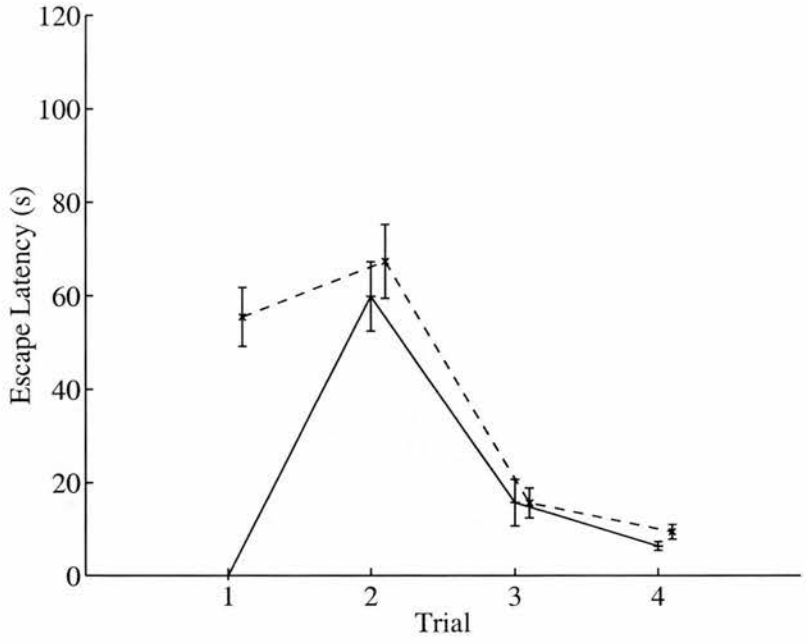


Figure 5.8: Variable platform control: mean latencies (N=12) over 2 blocks of testing, for swim days (dashed line) and placement days (solid line). Unlike the training phase, during this control phase, the platform occupied a novel position on trial 1 *and* a new, novel position on trial 2, in which it remained for the remaining trials of the day.

.001].

5.3.3 Discussion

The primary conclusion of this study is that placement is indeed sufficient to support one-trial learning, and this is reinforced by the fact that swim and placement conditions did not differ. The first prediction of the model of the previous chapter appears to have been upheld. The variable platform control was essential in order to attribute the improved performance to behaviour tied to the current platform location, and revealed that both swim and placement groups were profoundly affected by the changed platform location on trial 2.

As noted in the introduction, an important issue raised by previous studies is the extent to which improved performance after placement (*ie* an improvement between trials 1 and 2) is due to non-spatial factors or really due to the acquisition of something like a coordinate system, as predicted by the model of the previous chapter. With a larger number of rats, Whishaw (1991) found a difference between swim and placement. Did, however, that experiment succeed in eliciting the best possible performance from the placed rats? The best placed mean trial 2 latency in Whishaw (1991) was 21.9 seconds in a 1.46m diameter pool. By contrast, the mean trial 2 latency after placement in this experiment was much

smaller, at 15.3 seconds in a 2m diameter pool. As noted, this may be because of the way in which pretraining and training were not distinguished in Whishaw (1991); or it may be because only 2 swim trials were given per day in his experiment; or it may be that in fact the smaller pool failed to motivate the animals sufficiently. This latter point is emphasised by the difference between trial 1 latencies, when the location of the platform is not known to the rat: 65.2 seconds in this study, but only 40.6 seconds in Whishaw (1991). Therefore, the difference between swim and placement reported by Whishaw (1991) appears to be related more to the experimental protocol used than to the mechanisms involved in learning from placement.

5.4 Experiment 2: One-Trial Learning In The Presence of Barriers

5.4.1 Aims and Methods

The second experiment was designed to address directly the issue of one-trial learning in the presence of barriers. The delayed match-to-place (DMP) protocol was used to investigate transfer of spatial knowledge in a watermaze apparatus. However, the complexity of navigational demands was increased through the introduction of barriers. The use of barriers in a watermaze is a double-edged sword: while barriers add to the complexity of the required behaviour, they also introduce local cues that might render various elements of the navigational task rather more trivial than in the standard watermaze. Therefore the apparatus was designed to minimise the information provided by local cues alone about the appropriate action to take. Successful performance depended more on an understanding of the spatial relationship between the starting position, the barriers and the goal (platform) position, than upon any single one of these variables.

Overview of H-maze experiment

The aim of the experiment was to investigate transfer of navigational learning under changes in the goal position, and so a DMP protocol was used throughout. A barrier shaped like an 'H' was placed in the centre of the pool (figure 5.9). On each day, an escape platform occupied a novel position, but always within one of the two bays of the H barrier. Each rat started the first trial of each day from either the 'West' or 'East' position, facing one or other of the columns of the 'H'. On the second trial, the rat started from either 'North' or 'South', facing into one of the bays of the 'H', and had to choose whether to swim into the facing bay, or swim around the 'H'. In this way, small changes in the position of the goal would have drastic changes on the actions that were demanded, depending on which

side of the barrier the goal was located. Pilot experiments suggested that a choice of paths differing only in length provide insufficient motivation to choose the shorter, even in a watermaze. Therefore incorrect actions in the H-maze were designed to lead to a "dead-end" such that the rat had to then retrace its steps almost back to the starting position and take the correct action. Finally, to increase the negative effects of making an incorrect choice, the Atlantis Platform was used instead of a conventional solid platform (Spooner *et al*, 1994). This platform sits at the bottom of the pool and only rises to the surface if the rat swims within a circle of water above it, for a required length of time. The procedure has been shown previously to lead to more accurate searching by rats (Spooner *et al*, 1994).

Subjects

The subjects (N=12) were experimentally naive male Lister hooded rats.

Apparatus

The basic apparatus (watermaze pool, video camera and basic image analysis) were the same as in experiment 1.

The barrier was a single frame composed of three sheets of aluminium, in the shape of an 'H', as shown in figure 5.9. The centre-piece of the 'H' was 120cm. long, while the two outer pieces were 80cm. long. The barrier was arranged on top of bricks so as to be sufficiently high above the water level. It was essential that the head-on views of both bays of the barrier were exactly the same. The placement of rivets and supports were chosen to, as far as possible, support this similarity.

The Atlantis Platform was controlled by the Acorn application (written by R. Spooner). The 'zone' of the platform was defined as a circle of radius of 20cm. centred on the platform. The platform was automatically raised after the rat spent the required amount of time within this zone.

Procedure

Pretraining

Rats were handled for several days before experiencing the watermaze. The rats were pretrained using an adapted version of a regime that has been developed in the Edinburgh laboratory (Micheau, J., pers. comm.) for use with the Atlantis Platform, and which has proved to be very successful. The aim of the regime is to develop appropriate behaviour

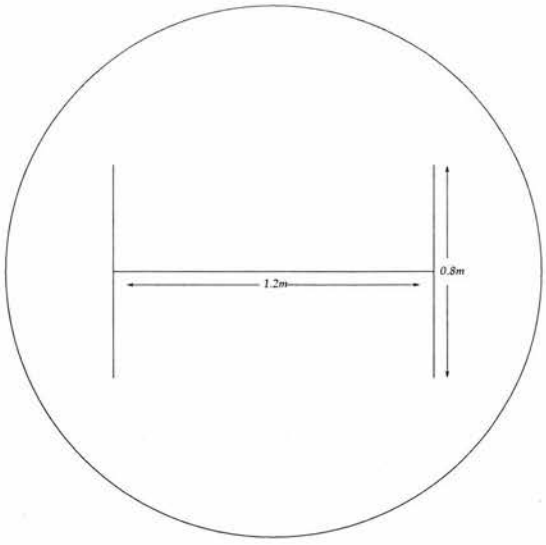


Figure 5.9: The dimensions of the H barrier relative to the watermaze. The diameter of the watermaze itself was 2m.

towards the Atlantis Platform. On successive days, rats are required to spend increasing amounts of time within the platform zone before the platform rises. However, a minimum amount of spatial learning is desired, hence the following steps are taken. 1. White curtains are drawn around the watermaze, obscuring the extramaze cues. 2. The platform position varies randomly throughout the maze *from trial to trial*. 3. A clear and brightly striped cylinder is hung immediately over the (hidden) platform position. The curtains probably do not completely obscure visual cues, and there are probably many other cues, *e.g.* auditory cues, available to the rat. However, the hope is that in combination with the other two measures, the rats' attention (and perhaps therefore learning) is drawn away from the allocentric spatial location of the platform, and towards the cue, and the Atlantis strategy.

Rats were split into two groups, 6 animals in each, and would be maintained in these groups throughout the entire experiment. The first group always encountered the barrier at one particular orientation, and will be referred to as the 'H' group. The other group always encountered the barrier at an orientation that was 90^o different from that of the first group, and will be referred to as the 'I' group. The reason for the two barrier orientation conditions was to control for the use of particular extramaze cues.

Each rat received 10 trials per day, for four days. On days 1 and 2, the barrier was absent, that is, rats swam in the open-field watermaze. On days 3 and 4, the barrier was in the pool. The required dwell-time, that is, the time for which a rat was required to stay within the platform zone before the platform rose, was increased from day to day: on days 1 and 2 it was 0.5s, on day 3 it was 1s, and on day 4 it was 2s, at which level it remained throughout the rest of the experiment. Before *each trial*, the platform (along with the overhanging cue) was placed in a new location unrelated to the previous trial's platform location. Trials

proceeded as in experiment 1 (section 5.3.1, procedure, pretraining), with the extra detail that if after 60s the rat had failed to raise the atlantis platform through dwelling, it raised automatically.

Training

Each rat received 5 trials per day, for 22 days. During training, the hanging cue was removed, and the curtains were drawn back and bunched together, allowing the rat to see the extramaze cues, and forming an additional extramaze cue. The basic procedure for beginning and ending a trial was the same as for the pretraining; the only differences were the choices of starting and platform position. Note also that the the procedure for the 'I' group was the same as for the 'H' group, but with everything (starting positions, platforms positions, barrier) rotated through 90^o (except for days 1 and 3 of training, for which platforms were located at -90^o relative to those in the 'H' group).

At the beginning of each day, the platform was placed in a strictly novel position, within one of the two bays of the barrier, and remained in the same position throughout the day. An example set of trials is shown in figure 5.10. On trial 1, each rat was started from either the 'west' or 'east' position. On trials 2 - 4, the rat started each trial from either the 'north' or 'south' position, facing into one of the bays of the barrier. The sequence of 'north' and 'south' starting positions, and the trial 1 starting position, and the side (north or south) of the barrier on which the platform was located, were all randomised, that is, they formed a counterbalanced, randomised sequence across all 22 days. However, from days 1 to and including day 12, the sequence of starting positions (with respect to the barrier) was the same for all rats, and hence all rats faced the same choice of action on trial 2. Then, from day 13 until day 22, the rats were split further into two groups (four groups of 3 rats each: two 'H' groups, and two 'I' groups). The sequence of starting positions of the rats in one group was the opposite of that in the other. Hence, on trial 2 of these days, half the rats were required to go around the barrier to get to the platform, and half were required to go straight into the bay in front of them.

Transfer Tests

Two sorts of transfer test were used, to investigate what the rats may have learned. Both tests involved only 2 trials. In one test (the MOVE test), rats experienced trial 1 to a novel platform position as during training (starting at 'west' or 'east'). After trial 1, and before trial 2, the barrier was *moved* so that the platform was now on the *opposite* side of the barrier from where it was on trial 1, although *the platform itself had not moved* (figure 5.11). Trial 2 then continued as during training (starting at 'north' or 'south'), but with the barrier in the new position. In the other test (the REMOVE test), trial 1 involved a novel platform position, as during training, but the procedure on this trial was very different. The barrier

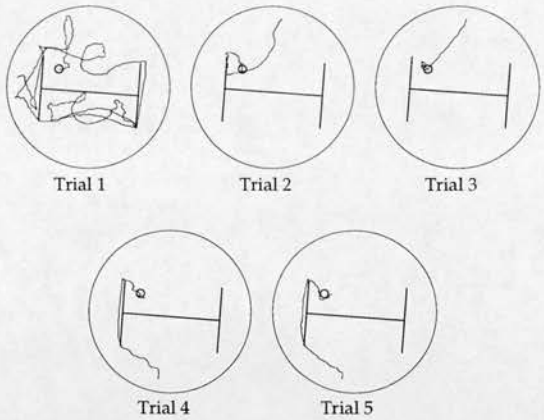


Figure 5.10: The paths taken by one of the rats on the 7th day of training. Trial 1 starts from the west position, and the rat searches for the platform which is in a novel position. Trial 2 starts from the north position, and the rat must choose whether to go around or go straight. In this case, the rat chooses to go straight, which is correct. Trial 3 also starts from the north. Trials 4 and 5 start from the south, and the rat chooses, again correctly, to go around.

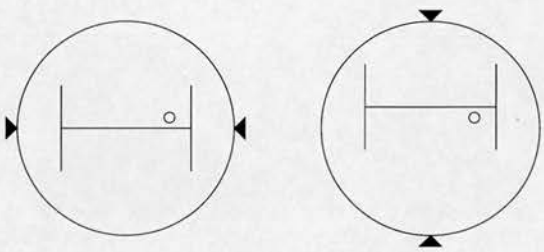


Figure 5.11: A schematic diagram of the MOVE transfer test: trial 1 begins at either the east or west position, and the platform occupies a novel location. Trial 2 begins at either north or south, however although the platform remains unmoved, the barrier has been moved such that the correct choice of whether to go straight to the platform, or go around the barrier, is reversed.

was removed. Rats were taken from the cage at the side of the room, and dipped, while being held, into the water, and then quickly placed onto the platform (which was raised) and left there for 1 minute. The trial ended with them being collected and placed again at the side of the room. Subsequently the barrier was put back, in its training position, and trial 2 was conducted as in training (starting at 'north' or 'south'). Note that although the barrier was put back, *the platform was not moved* between trial 1 and trial 2.

The tests took place in a sequence designed to minimise the adverse effects of the changes in procedure. The groups were altered: both 'H' and 'I' groups were split into two groups of 3, such that the range of performance in each group was roughly similar. Note that the resulting bisections were not the same as those made during training. Rats in the first subgroup, from both 'H' and 'I' groups, underwent a MOVE test on day 1, a normal (training) day on day 2, a REMOVE test on day 3, and a normal day (two trials only) on day 4. Rats in the second subgroup, from both 'H' and 'I' groups, underwent a normal day on day 1

(two trials only), a REMOVE test on day 2, a normal day on day 3 and a MOVE test on day 4.

Data Collection and Analysis

The majority of the analysis was carried out by transferring the rat position data from the Acorn platform into UNIX files, for processing on a Sun workstation, using software written by the author. Firstly, the data files were processed to remove tracking errors (errors made when the image analyser "lost" the rat during a trial). Secondly, various kinds of analyses were performed on the path data.

Tracking Errors

The extraction of position data from the video signal by the image analyser relies upon a head-tracking system which picks out, from a background of semi-opaque, latex-filled water, the black mark characteristic of the back of the head of Lister hooded rats. In the standard apparatus tracking errors occasionally occur - the system loses track of the rat for short intervals within a trial. The barrier exacerbates the problem, since a rat's head becomes occluded to the camera when the rat passes close to the barrier, and also because the tracking system occasionally jumps to some part of the barrier itself. The interpolation provided by the Acorn application was adequate for conventional watermaze experiments, but too crude to support meaningful estimates of path length when the barrier was present. Therefore, a new program was developed for the appropriate detection and handling of tracking errors. Running the program on a given path identifies errors from 4 possible types, as follows: (1) if any point is outside the pool; (2) if the first point is within 30cm of the centre of the pool; (3) if the rat jumps faster than a maximum speed (chosen by the experimenter, if necessary on a trial-by-trial basis); (4) if the current position makes a jump, the subsequent path is only classed as veridical if it is not followed soon after by another jump (the time between jumps can also be set, if necessary on a trial-by-trial basis). The last error was found to be useful because most erroneous jumps, *e.g.* to some part of the barrier, were followed soon after by a jump back to the rat. By contrast, if the rat was obscured for some period, *e.g.* behind the barrier, the jump would be an appropriate interpolation of the data, and would not usually be followed by another jump. Examples of paths before and after this clean-up operation are shown in figure 5.12.

Analyses

The clean path length data achieved by the program allowed path lengths to be calculated, and from that the measure finally used: the path length minus the shortest possible path length from start position to platform, for each rat for each trial, referred to as "path length

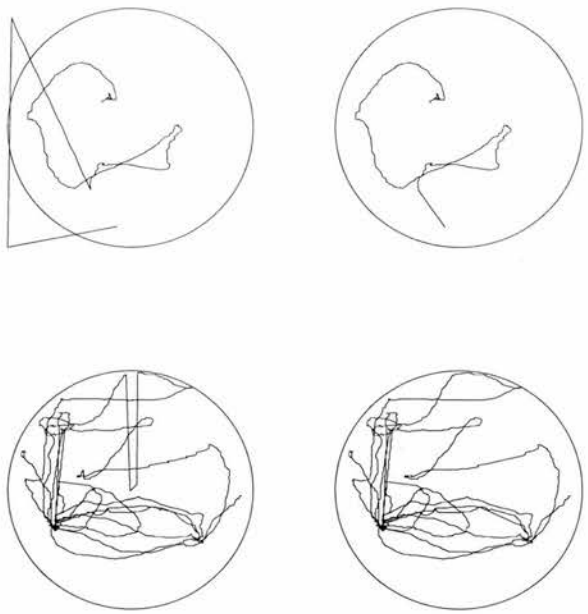


Figure 5.12: Two examples of the process of cleaning up of tracking errors. The paths on the left are output by the Watermaze program. The tracking errors are clearly visible. The paths on the right have been cleaned up. For a description of the principles upon which this process is based, see text.

minus optimal".

A new heading analysis was developed to examine the navigational strategy of the animals far from the platform. The question at stake was whether or not the performance of the animals truly reflected the optimality of action choices far from the platform - as, for example, when they had to choose to go straight into the barrier or go round, or whether the performance reflected a rather more local strategy of knowing what to do when close to the platform, and searching in a rather less well directed manner further away from it. This question is critical: did the rats show one-trial solving of a non-trivial temporal credit assignment problem in navigation, or not? First, an instantaneous heading direction as a function of time was extracted by smoothing the data, and calculating a heading between pairs of successive smoothed points. Second, this data was used as the basis for a type of zone analysis.

The instantaneous heading was, like all derivatives calculated from real data, likely to be extremely noisy, and this could cause serious problems for an analysis based upon this data. For example, it is easy to imagine that a path leading clearly in a particular direction could appear to be composed of two different directions due merely to this noisy derivative. Furthermore, the sampling of the image analyser itself demands a certain degree of smoothing. This was achieved by convolving the path data with a gaussian function, in the manner of a low-pass filter. The particular choice of parameter is quite important – too much smoothing might lead one to conclude that all paths were heading in the direction of interest, whereas

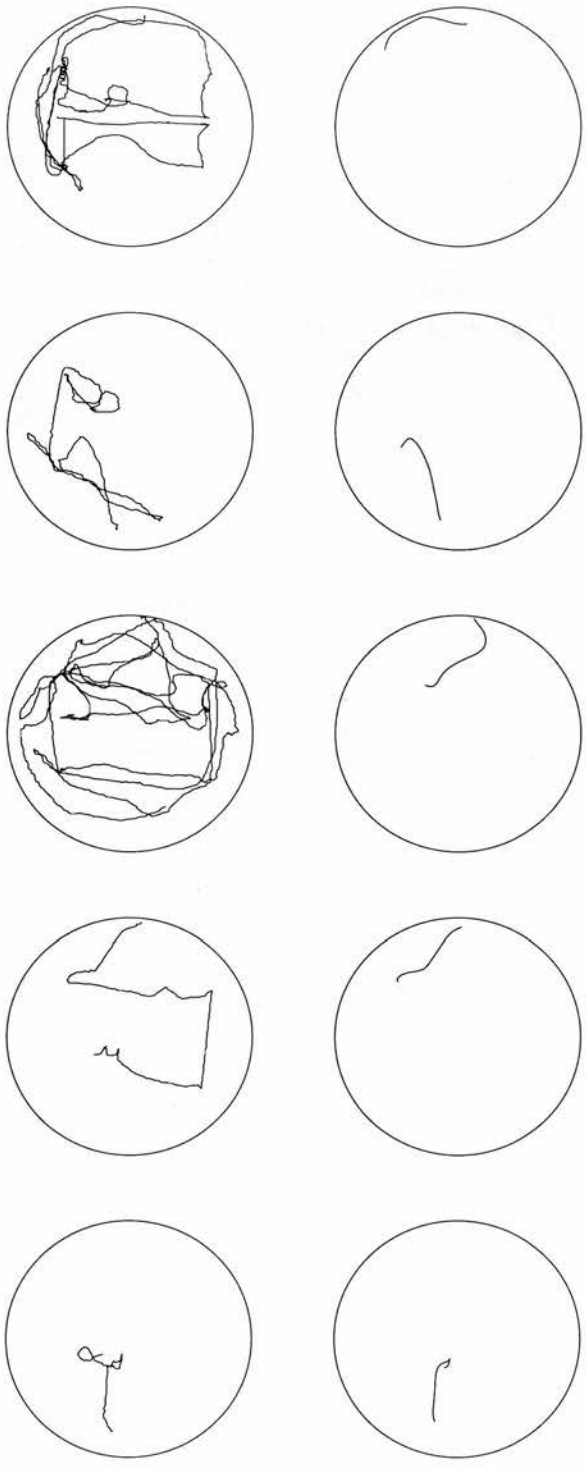


Figure 5.13: Examples of smoothed paths. The figures on the left show path data cleaned of tracking errors, but not yet smoothed. The right figure shows the smoothed versions, over the first 5 seconds only. This was all that was used in the zone analysis (see text). The smoothed paths clearly capture the heading of the animal, while removing sampling noise. The smoothing is narrow enough to allow windy paths to remain windy.

too little would leave the noise problem unchecked. Figure 5.13 shows smoothed data for several paths.

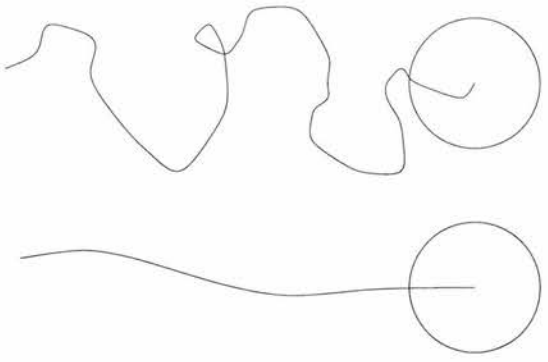


Figure 5.14: The amount of time spent heading towards a zone of interest reveals something about the directness of the navigational path. The top path is windy and would lead to less time spent heading towards the zone than the lower path which is direct. Note that a similar conclusion can be reached by considering only an early part of the path; unlike zone analysis based on *position*, it is not necessary for the rat to have reached the zone to say something about its navigational performance.

Conventionally, zone analysis is use to determine the extent to which a rat searches in a given place, *eg* Steele and Morris (1999) found it useful to measure, for each trial, how much time the rat spent in the zone around the platform of interest, as a fraction of the time spent in any of the zones around each of the platform positions used throughout the experiment. For any single trial, such an analysis can reveal if the rats had a preference for occupying a particular spatial location. Across a set of days, where the experimental design manipulates which platform is of interest, such an analysis can reveal if indeed the manipulated factor (*eg* previous day's platform position) is controlling the rat's search.

In contrast to the conventional approach, based on the *position* of the rat, a zone analysis was developed based upon the rat's *heading*. The principal advantage of this approach is that it enables the evaluation of a rat's preference for a given location while the rat is still far from the location, a feature particularly useful for the question of one-trial learning. Furthermore, the potential difference between the behaviours of navigating to a location and recognising a location once there, possibly exacerbated by the dwelling required by the Atlantis platform, is ignored by the conventional zone analysis. Zone analysis by heading is potentially effective in distinguishing between good and poor navigational trajectories towards a location, as demonstrated by figure 5.14.

Zone analyses by heading were carried out on the first 5s of trial 2 only. One analysis was carried out with a circular 20cm diameter zone centered on the current platform location, referred to as the *platform analysis* (figure 5.15). A second analysis was carried out with the same size of zone centered on one of the four corners of the barrier, chosen as follows, and referred to as the *barrier analysis* (figure 5.15). First, for each trial the corner was one

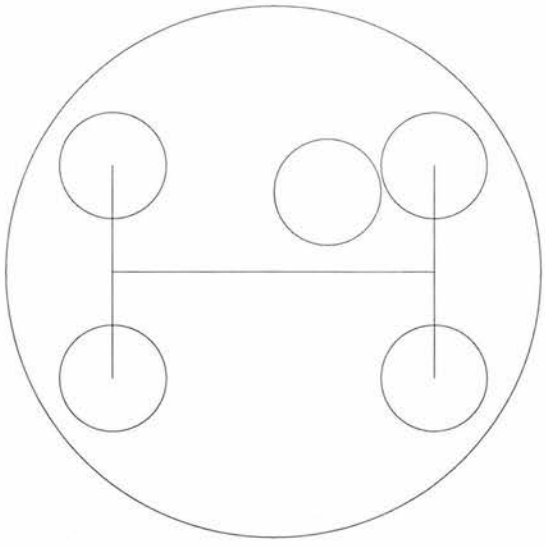


Figure 5.15: Zones were centred on either the platform (platform analysis), or on one of the four corners of the barrier (barrier analysis). For the barrier analysis, the choice of zone was made, for each trial, on the basis of which side of the barrier the rat was on, and which route it took around the barrier.

of the two on the side of the barrier facing the rat. Second, the choice of which of the two corners was to be the zone was defined as that corner by which the rat eventually traversed to the opposite side of the barrier, *ie* the zone was defined by the rats' own movements, and so this was not a measure of whether rats chose the optimal of the two possible trajectories around the barrier – they almost certainly did not as there was little motivation offered by the experiment for them to do so. In those cases when no such corner was defined, *eg* when the rat correctly went straight, the zone producing the largest value was used.

5.4.2 Results

Pretraining

Mean latencies for pretraining reveal an improvement over the very first few trials, presumably due to the animals learning various strategic aspects of the task, with little improvement thereafter. The addition of the barrier on day 3 led to a noticeable increase in latencies. The important feature of the pretraining, however, is that rats successfully learned how to dwell.

Training

During training, most rats developed a good navigational strategy, *ie* they tended to search the maze thoroughly on trial 1, and move directly to the platform on trials 2-5. However, two of the rats developed a strong tendency to swim directly to one of the corners of the

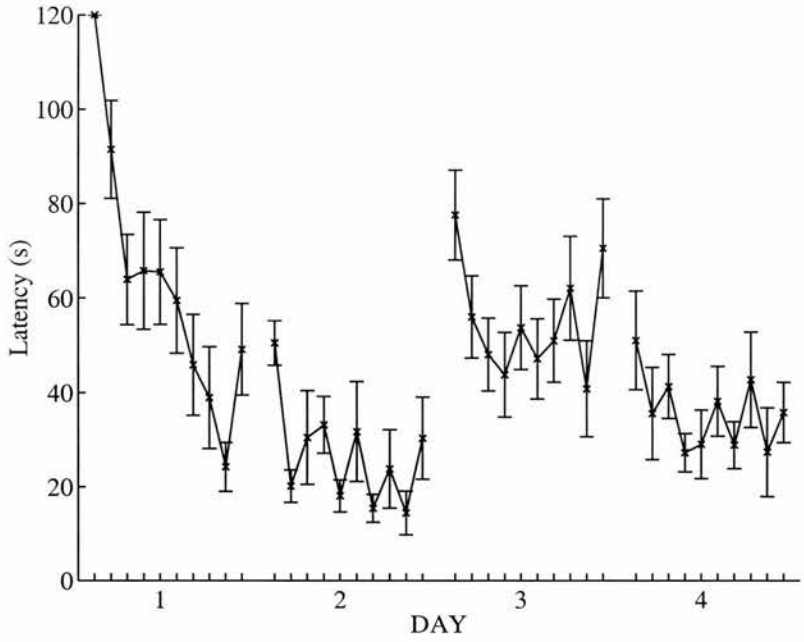


Figure 5.16: Pretraining: mean latencies for pretraining, during which rats become accustomed to dwelling for the Atlantis Platform for progressively increasing times (day 1: 0.5s; day 2: 0.5s; day 3: 1s; day 4: 2s). The barrier was absent for days 1 and 2, and present for days 3 and 4.

maze and hang there for up to 120 seconds, *ie* for the duration of the trial. Although it was possible to train these animals not to hang by removing them from the barrier whenever they attempted to do so, nevertheless the hanging lasted for several days, and so it was necessary to remove these animals from subsequent analysis. Unfortunately, both animals were from the 'I' group; fortunately each animal was in a different group with respect to 'straight' and 'round' choices (from day 13 onwards).

The measure "path length minus optimal" reveals the development of one-trial learning, in a similar manner to previous experiments (figures 5.17 and 5.18). The key issue, however, is whether there is a difference using this measure between having to go straight and having to go round: optimal performance should show no difference, but the use of different strategies in each case might.

Following the procedure of experiment 1, and as suggested by the model of the previous chapter, days 1-6 of training were regarded as a pretraining period. The measure "path length minus optimal" was compared for trial 2 only, across days 7-22. The mean for "straight" trials over this period was 2.91 $\pm$ .56, and for "round" trials, 2.92 $\pm$ .54. A repeated measures ANOVA of this data with "straight"/"round" as a within subjects factor revealed for swim days no significant difference whatsoever [$F(1,11)<1$; $p > .9$].

Given that path lengths were still somewhat longer than the optimal paths, there is clearly

a lot of variability in the behaviour of the rats. The possibility remains that rats do not or can not solve a non-trivial credit assignment problem, in one trial. In order to sufficiently demonstrate such an ability, rats must be seen to engage in differential and optimal behaviour dependent upon whether or not the platform is on the same side of the barrier as the starting position, and moreover, this behaviour must be clear on the second trial to the platform position. Therefore a zone heading analysis was performed.

Heading data was acquired by smoothing the paths and taking the first derivative. Zone analysis was conducted on this heading data for the first 5 seconds of trial 2 for days 7-22, only. Zones were defined for the platform position for each day, and for one of the corners of the barrier. The percentage time spent heading towards the zones is shown in figure 5.19. For the platform zone, a repeated measures ANOVA performed with "same side"/"opposite side" as a within subjects factor revealed that animals starting from the same side of the barrier as the platform spent significantly more time heading directly towards the platform (in the first 5 seconds of trial 2) than animals starting from the opposite side [$F(1,9)=83.213$; $p < .001$]. For the barrier zone, a repeated measures ANOVA with "same side"/"opposite side" as a within subjects factor revealed that animals starting from the opposite side of the barrier to the platform spent significantly more time heading towards the corner of the barrier (in the first 5 seconds of trial 2) than animals starting from the same side [$F(1,9)=8.409$; $p < .02$].

Transfer Tests

On trial 2 of the MOVE test, rats made the choice of going straight or round that would have been correct if the barrier remained as it was on trial 1 (figures 5.20 and 5.21). Perhaps more surprisingly, however, having swum to what was (on trial 2) the wrong side of the barrier, the rats clearly butted head-first into the barrier, in an apparent attempt to move through it to the platform. In other words, having taken the initial decision to go straight or round, they did not then rely solely on the local cue information provided by the barrier to direct their search strategy.

By contrast, the REMOVE test was less revealing as generally the rats appeared confused by the test, and after the placement failed to search directly for the platform (data not shown). This can presumably be attributed to the effect of removing the barrier on trial 1.

5.4.3 Discussion

In the presence of barriers, rats can develop one-trial learning abilities comparable to those shown previously in the same environment without a barrier. The particular decision forced

on the rats by the barrier used was to choose whether to head in to the facing bay of the barrier or to go around. In support of the claim that a directed navigational strategy is used, the performance on trial 2 was shown to be the same (when compared to an optimal path in each case) whether a rat started from the same or opposite side of the barrier as the platform. Moreover, on second trials in which an animal had to go straight, more time (out of just the first 5 seconds of behaviour) was spent heading towards the platform than when the animal had to go around. Likewise, on second trials in which an animal had to go around the barrier, more time was spent heading towards the corners of the barrier than when the animal had to go straight.

Has the introduction of the barrier made the navigation problem trivial through the introduction of local cues? First, the correct choice of whether to go straight or around was dependent on the locations of the platform, starting position and barrier all together, and so arguably the local cue information might have helped the rats decide where to search once on the correct side of the barrier, but not *which* side was correct, or what to do to get there. Second, local cues may not have been relied upon even to this extent, since in the MOVE transfer test, having made a choice to go straight or around, rats generally proceeded to head straight for the platform location as defined allocentrically, rather than by the barrier.

The conclusion, then, is that the model of the previous chapter is incomplete to the extent that it cannot account for the one-trial learning abilities of rats in the presence of barriers, using as it does a single, environment-wide coordinate system. However, the benefits of a coordinate system for navigation in an open space remain compelling, and may explain the way rats bump headlong into the barrier on the MOVE transfer test. The critical extra facility required according to this viewpoint is that rats not only know coordinates for an environment, but where in an environment those coordinates are useful. Such a viewpoint implies that rats learn about the structure of the environment, and it is this hypothesis with which the next chapter is concerned.

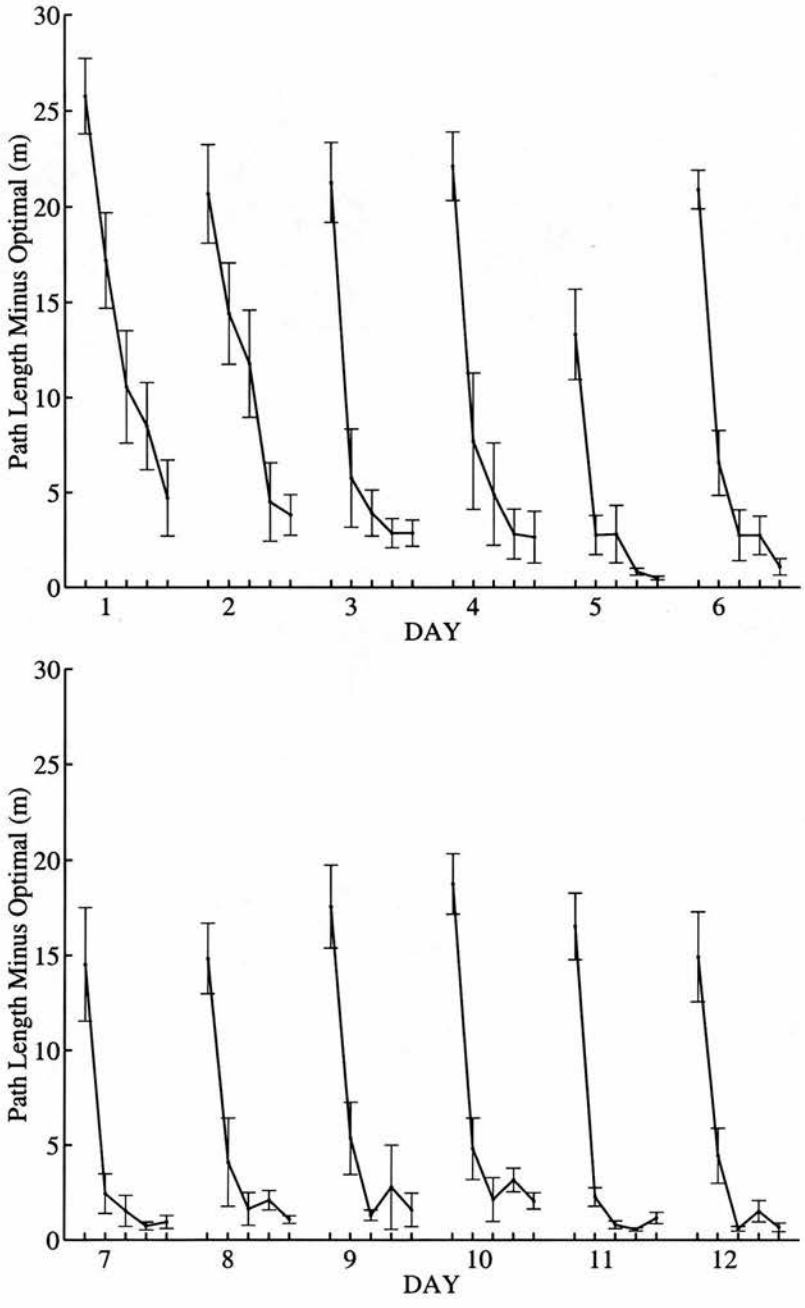


Figure 5.17: The means and standard errors across all 12 rats for the measure "path length minus optimal" (see text), for the first 12 days of training.

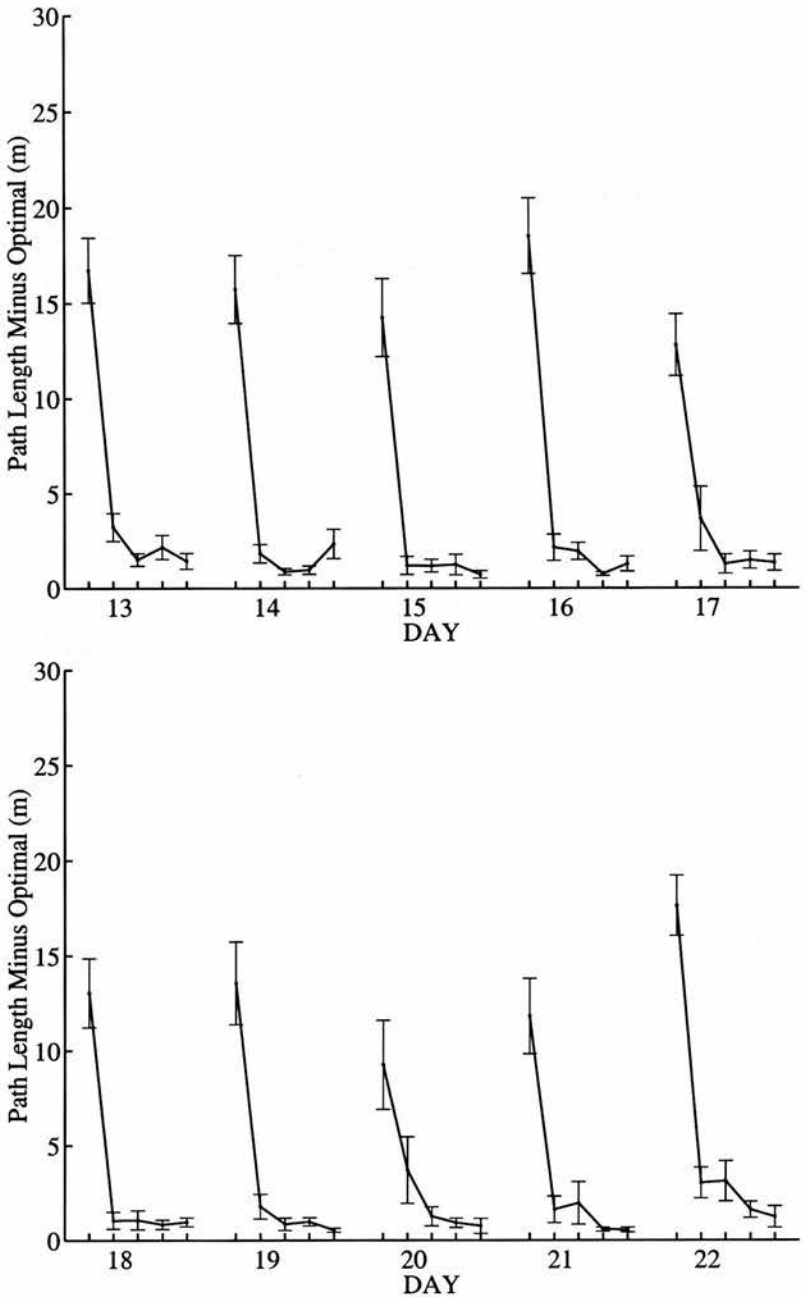


Figure 5.18: The means and standard errors across all 12 rats for the measure "path length minus optimal" (see text), for days 13 to 22 of training.

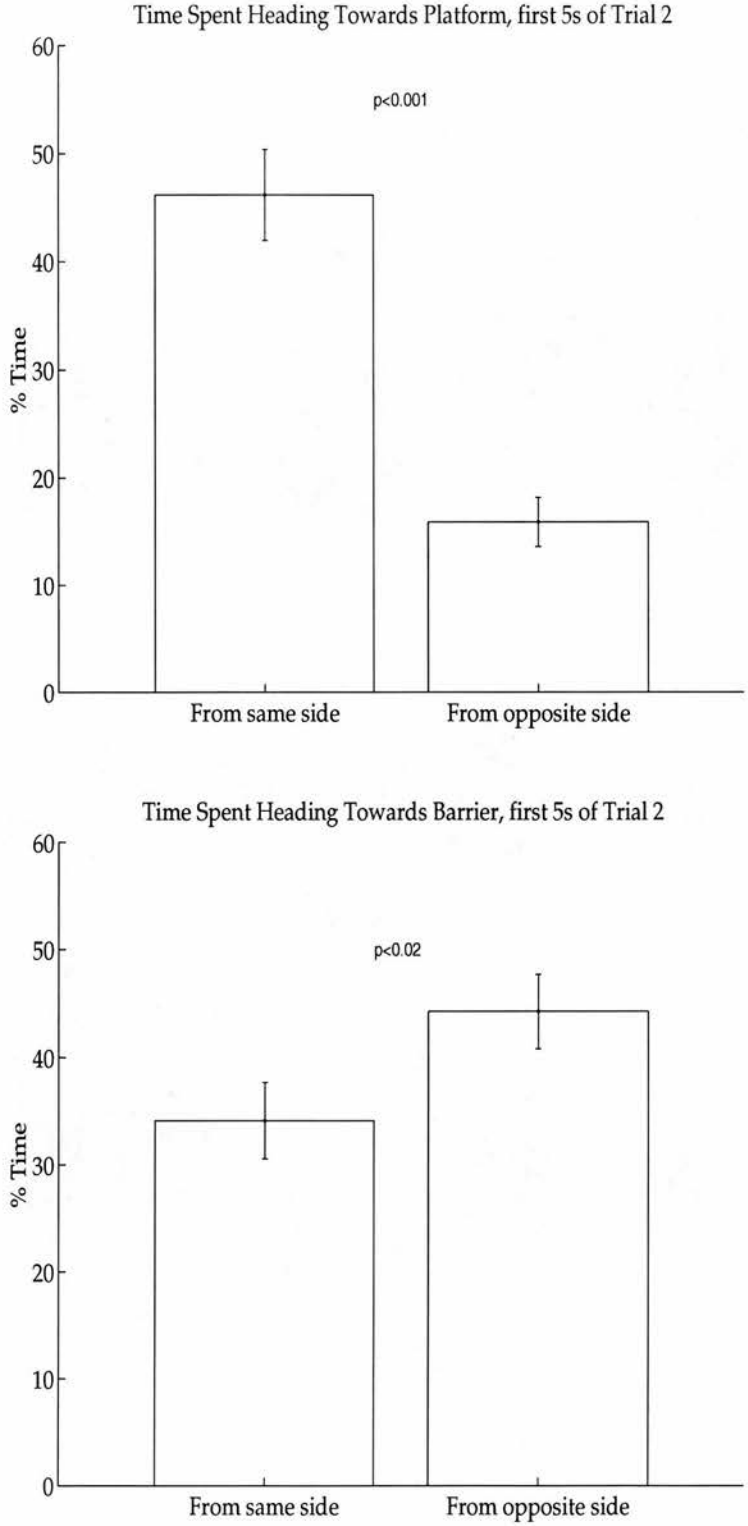


Figure 5.19: Zone analysis, the first 5 seconds of trial 2 only: (1) the top figure shows the percentage time spent heading towards a zone centered on the current platform position, from a starting position on the same or opposite side of the barrier as the platform; (2) the bottom figure shows the percentage time spent heading towards a zone centered on the near side corner of the barrier by which the animal crossed round to the other side (or to which the animal spent most time heading), from a starting position on the same or opposite side of the barrier as the platform.

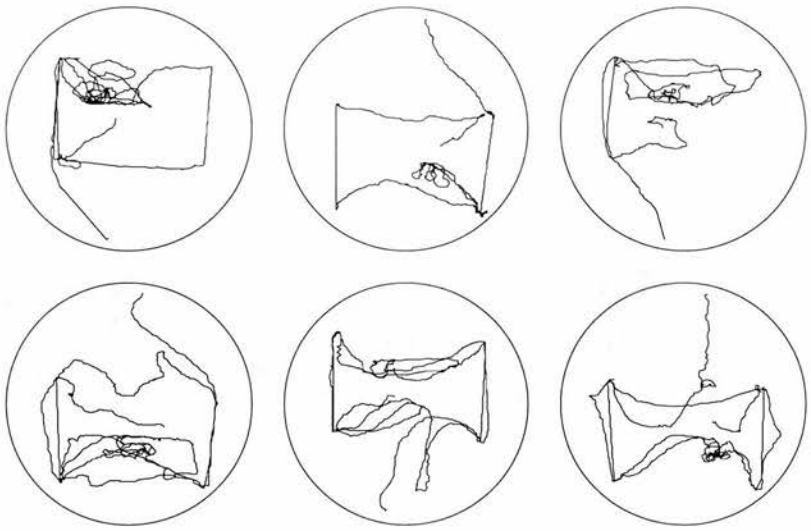


Figure 5.20: Transfer test: the paths taken on trial 2 of the MOVE transfer test, in which the correct action with respect to the position of the barrier on trial 1 was to go round. Rats initially went round, and then bumped into the barrier near to the platform.

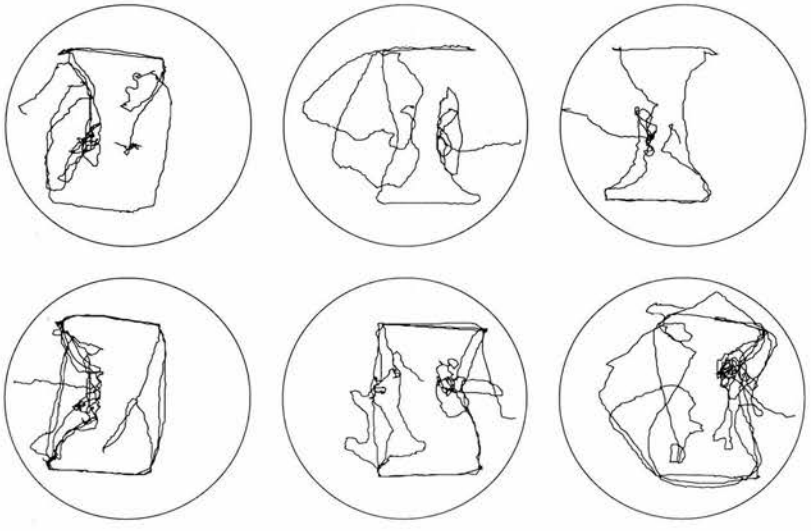


Figure 5.21: Transfer test: the paths taken on trial 2 of the MOVE transfer test, in which the correct action with respect to the position of the barrier on trial 1 was to go straight. Rats initially went straight, and then bumped into the barrier near to the platform.

Chapter 6

Using Unsupervised Learning To Find Structure in Value Functions

6.1 Introduction

The conclusion of the previous chapter has been that rats show goal generalisation in the presence of barriers, and therefore a global coordinate system by itself is insufficient to account for the rats' performance. This provides the motivation for this chapter: to find a way of supporting goal generalisation even in the presence of barriers. However, we do not want to lose the neurally plausible aspects of the model considered so far, *eg* by retreating to more explicit dynamic programming methods, such as learning complete transition functions and having to compute optimal policies offline.

A problem related to generalisation is that methods based on dynamic programming scale notoriously poorly with the number of states, and this is no less true of model-free reinforcement learning methods. The relationship to generalisation is clear if one considers incorporating the difference between successive tasks as an extra state variable. For example, a "multiple-goals MDP" is one in which both the current location and the location of the goal define state, such that an environment with, say, N discrete locations gives rise to a multiple-goals MDP with N^2 states. Goal generalisation in this case means choosing actions optimally in each of the N^2 states, without having to learn about each of the N^2 states separately.

6.1.1 Two Design Principles For Solving Large MDPs

A natural way of dealing with large MDPs is to try to reduce the whole MDP into a number of sub-problems. The following discussion focuses on navigation, and identifies two key, though loosely defined, design principles:

Hierarchical Principle

Far away from a goal, specification of sub-problems is likely to be similar for nearby states and, symmetrically, for nearby goals. Moreover, the number of states and goals for which similarity holds is likely to increase the further apart the states and goals are. It follows that sub-problems can be organised in a spatial hierarchy. Examples: moving within a room to a goal may involve specifying a different action for every location, whereas moving to goals in another room may always involve moving through a doorway, *ie* for the more distant goals, there is a clear sub-problem. Alternatively, the H-maze problem of the previous chapter may require all-to-all, coordinate navigation if the current location is on the same side of the barrier as the goal, but if it is on the other side, a simpler "go around" action can be specified. There is a natural hierarchy of imprecision.

Structural Principle

The next issue is then how to choose which states should be lumped together. The structural principle says that the most effective segmentation of space is one which obeys the underlying structure of the world. Example: in the H-maze problem, the barrier defines at least two fragments. A segmentation of states which doesn't take the barrier into account is unlikely to be as effective. One implication of this principle is that, in general, structure needs to be learned.

These principles can in theory be separated: hierarchies are possible that ignore underlying structure, and non-hierarchical segmentations of space are possible that do not. However the principles are in practice likely to be inter-dependent: (1) underlying structure may be naturally hierarchical, *eg* sides of barriers within rooms, rooms within buildings, buildings within streets; (2) the hierarchical principle applied without regard for structure is unlikely to be succesful.

6.1.2 Applying The Design Principles To RL

Hierarchical structures and forms of abstraction have recently been the focus of substantial work in RL, while less work has focused on learning underlying structure (Watkins, 1989; Singh, 1992a, 1992b; Dayan & Hinton, 1993; Dayan, pers. comm.; Kaelbling, 1993; Sutton, 1995; Thrun & Schwartz, 1995; Dietterich, 1998; Hauskrecht, 1998; Hauskrecht, Meuleau, Boutilier, Kaelbling & Dean, 1998; Parr, 1998; Parr & Russell, 1998; Precup & Sutton, 1998; Precup, Sutton & Singh, 1998; Sutton, Precup & Singh, 1998; Moore, Baird & Kaelbling, 1998). The motivation for the present work is best captured in relation to one application in particular called Feudal Learning (FL; Dayan and Hinton, 1993; Dayan, pers. comm.).

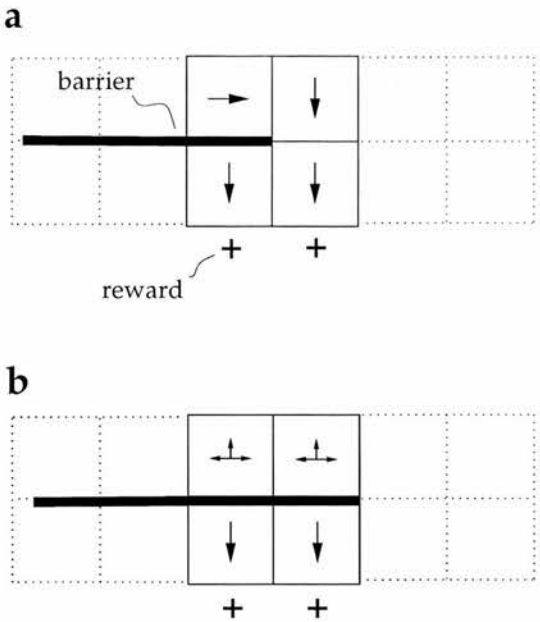


Figure 6.1: The effect of a single barrier on feudal Q learning. (a) Four vassals have to learn the command "S" which the master rewards as any transition that leaves the region by the southern edge. Even with the barrier shown, the vassals can learn appropriate actions, as indicated by the arrows. (b) A differently shaped barrier makes the rewarded transitions impossible for sequences beginning in either of the upper two states. At best, action choices in these states will be random, as shown by the arrows. However, these choices are sub-optimal with respect to the rest of the MDP outside of the region (sub-regions shown with dotted lines).

Inspired by a medieval feudal fiefdom, FL applies to a navigational MDP a predefined hierarchy of control. FL effects a deterministic, non-overlapping partitioning of the state space into regions, each of which has a controller associated with it. Each region in turn is split into sub-regions, each with a sub-controller associated with it, and so on for as deep a hierarchy as required. Communication between controllers only occurs between a controller and the one immediately higher controller whose territory it occupies, *ie* between master and slave.

The master issues a command to its slaves and rewards them upon completion, in a very particular way. When the agent is located within a master's territory, that master can issue its slaves with a command such as "move north". It's slaves, however, do not know *a priori* what this means for them, so they have to learn what sub-commands, or, if they are bottom-level controllers, what primitive actions, they in turn need to specify. The master *rewards* the slaves if the agent moves outside of its (the master's) territory, within a certain maximum allowed time, and by the desired "north" boundary. Other outcomes are *punished*. Hence each master creates a sub-MDP for its slaves.

Simulation results demonstrate dramatic increases in learning rate compared to conventional RL, *eg* feudal Q-learning applied to a multiple-goals navigational MDP easily beats

conventional Q-learning (Dayan and Hinton, 1993; Dayan, pers. comm.). However, there are two problems with FL. The first follows from the fact that FL ignores structure. Just one barrier added to the (completely unobstructed) environment in the simulations could, if inconveniently placed, lead to highly sub-optimal performance. While one nice aspect of FL is that a master does *not* need to know everything about barriers within its territory, nevertheless a badly placed barrier can induce partial observability, so that a master cannot specify a command appropriately for all the slaves in its territory (figure 6.1).

The second problem with FL is that its rigidly hierarchical control structure is unlikely to be neurally plausible. By contrast, the actor-critic method examined in previous chapters specifies control within a single selection and evaluation mechanism; where separate control modules were considered, such as the coordinate controller alongside primitive actions, they were made to compete within the same, simple selection system.

What the above discussion of FL suggests is that the hierarchical principle might be applied very effectively if combined with the structural principle, *ie* learning about the underlying spatial structure of the environment. Therefore the main focus of the present work is an attempt to learn this structure by clustering states on the basis of values (see section 6.1.3). A secondary focus is on a way of using hierarchical structure for navigational learning without recourse to the rigid hierarchical control mechanisms of FL (see section 6.1.4).

6.1.3 Learning Structure By Clustering

We wish to cluster states together, and generally in a hierarchical manner, on the basis of their association with sub-problems defined by the navigation problem in hand. Of course, the nature of these sub-problems is unknown. Therefore, some surrogate feature is required on the basis of which states may be clustered.

In selecting such a feature, the key design principles are important. One choice might be to cluster states according to the similarity of *optimal actions* taken in states. Across a set of tasks defined by different goals, this may indeed be a fruitful strategy. However, the range of actions, which might be as few in number as four, is arguably somewhat restrictive.

A second option is to use *optimal values* appropriate to an MDP defined by one goal, or even the values defined by several different goal locations. Despite being more abstract that actions as a basis for clustering, nevertheless values obey both the structural and hierarchical principles. The first is simple to see in the case of a barrier dissecting some area, for which values are much more likely to be similar for a set of states on the same side of the barrier, than for a set of states drawn from either side. The second principle is evident in that values may be clustered at a variety of scales, *ie* an hierarchy can be imposed upon

values. Moreover, values span a range somewhat related to the size of an environment, and so may provide a sufficiently rich data set for clustering. Therefore, the hypothesis of the present work is that clustering values is an appropriate way to determine the underlying structure of an environment, and most of this chapter is concerned with examining this claim.

6.1.4 Using Structure To Aid Learning

Having created a hierarchical decomposition of the navigational MDP which also obeys underlying structure, the natural question is to ask what one intends to do with it. A natural answer is to use the hierarchical decomposition to define a hierarchy of spatially localised, separate component MDPs, just as in feudal learning (FL). However, a simpler possibility is investigated in this thesis.

FL works by making masters teach their slaves by rewarding changes in state that are correct when judged at the master's spatial level, *eg* leaving the master's domain by the desired boundary. However, exactly the same effect might be achieved by more simply allowing both master's and slaves' domains to acquire value in a conventional RL manner. That is, by allowing *state representation at multiple, concurrent resolutions*, resulting borders between values at different hierarchical levels should automatically mimic the very rewards which FL has to specify explicitly. Ultimately, an appropriate representation might prove as beneficial for learning as a complicated, hierarchical control and evaluation structure.

Therefore, this thesis considers learning hierarchical structure and then using this structure to produce a multi-resolution representation of state, in which high-level fragments and rooms *augment* low-level states. Computationally, however, the representation could not be simpler: high-level representations will simply report on the occupancy of the associated domain by the agent. This state representation will be used with TD-learning within an actor-critic architecture of the sort already extensively considered (section 3.4.6; chapter 4).

6.1.5 Chapter Layout

Section 6.2 describes the families of multiple-goals navigation problems used to illustrate the approach; section 6.3 then introduces unsupervised learning as a suitable method for extracting structural information; sections 6.4, 6.5 and 6.6 discuss the different probabilistic models used to capture the structure of value functions; and section 6.7 describes results on the utility of the decompositions for actor-critic learning. The significance of the approach and results is discussed in section 6.8.

A B C

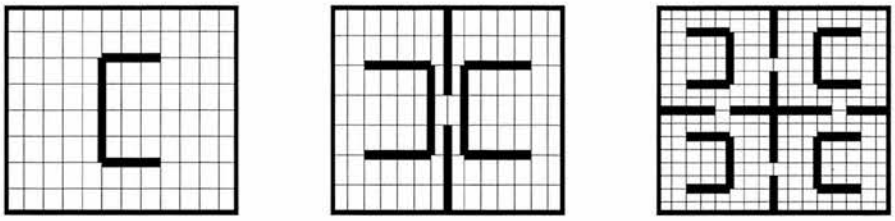


Figure 6.2: Simple mazes. The thin grid lines outline the states the agent can occupy; the thick lines show impassable walls. Manhattan moves (North, South, East and West) are permitted, and are deterministically successful.

6.2 Multiple-Goals MDPs

For convenience, a slight retreat is made from the continuous location space of chapter 4, the aim at this stage being to establish the plausibility of the basic ideas. Figure 6.2 shows three example environments which are used. Each example involves a simple discrete grid-world (with 96, 98 and 256 states), although their structures are rather different. Manhattan moves (*ie* non-diagonal straight moves only) are possible everywhere except at the boundaries and across the barriers (thick lines) and, again for convenience, are taken to be deterministic.

In a multiple-goals MDP, any of the states can be a goal, therefore we are interested in solving many similar problems. The differences between them come in the reward structure, and also to a lesser extent in the dynamics of what happens at the goal, which is absorbing. In contrast to the approach of previous chapters, discounting is not used but instead costs $R(s,a)$ of 1 are associated deterministically with each move from state s using action a to non-goal states, and costs of 0 with moves to goal states. Figure 6.3 shows optimal value functions for the three mazes for two different locations of the goal. The optimal value function for state s when the goal is g is defined as

$$V^g(s) = \min_a \left\{ R(s,a) + \sum_{s'} \mathcal{P}(s'|s,a) V^g(s') \right\}$$

where $\mathcal{P}(s'|s,a)$ is the transition matrix (consisting of 0s and 1s) reporting the consequence of applying action a at state s (strictly $\mathcal{P}^g(s'|s,a)$, since the process absorbs at the goal g). The optimal value function, in conjunction with $R(s,a)$ and $\mathcal{P}(s'|s,a)$, can be used to determine optimal actions at each state.

Figure 6.3 shows clearly how the optimal value functions reveal the underlying structure of the environments. For each pair of value functions, even though the values within each segment of the mazes are quite different, the underlying discontinuities across the barriers are the same. This is why unsupervised learning of the value functions should extract

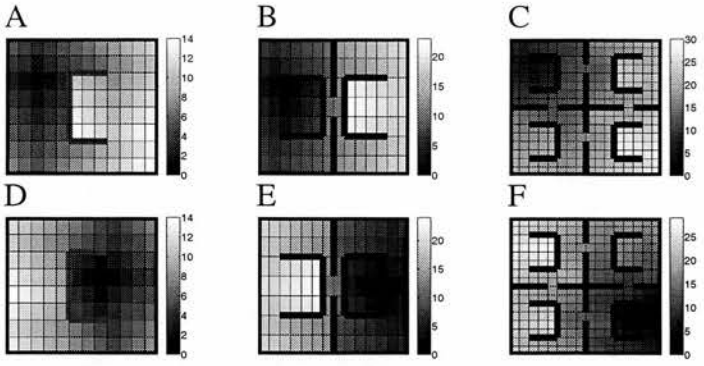


Figure 6.3: Grayscale value functions for the mazes in figure 6.2 for two different locations in each case. Note the different scales for each of the mazes.

appropriate structural decompositions.

6.3 Unsupervised Learning

Unsupervised learning (UL) is the process of learning about structure in a set of input data, without reference to any supervisory information. In reinforcement learning, the information the environment provides for learning is considered to be extremely sparse; for unsupervised learning it is simply non-existent. UL is relevant to learning situations in which a very large amount of unlabelled input data is required to be processed, and so is often applied to sensory systems such as both artificial and neural vision. In the present work, UL is applied to navigational, multiple-goal MDPs in which there are a very large number of states. It is fairly unusual to combine UL with RL, however, despite the fact that both types of learning are likely to occur in the brain.

Evidently in the absence of supervisory information, a learning task must be constructed. One method, *probability density estimation*, involves attempting to account for the input data in terms of a parametrised probabilistic model, sometimes referred to as a *generative* model. The method is appealing because learning rules have a precise, probabilistic interpretation but also because of the intuitive appropriateness of attempting to build an underlying model to account for the data. However, density estimation is only a surrogate for the true though less quantifiable aim of learning good representations (*eg* Hinton and Ghahramani, 1997).

The input data upon which unsupervised learning will be performed consists of the set of *optimal* value functions $\mathcal{D} = \{\mathbf{V}^g\}$ for a collection of goals g. For a parametrised probabilistic model, we wish to find the best parameters with which to explain the data, *ie*

find the parameters θ^* that maximise $\mathcal{P}[\theta|\mathcal{D}]$. However, Bayes rule informs that:

$$\arg\max_{\theta}\{\mathcal{P}[\theta|\mathcal{D}]\} = \arg\max_{\theta}\left\{\frac{\mathcal{P}(\mathcal{D}|\theta)\mathcal{P}(\theta)}{\mathcal{P}(\mathcal{D})}\right\} \quad (6.1)$$

$$= \arg\max_{\theta}\{\mathcal{P}(\mathcal{D}|\theta)\mathcal{P}(\theta)\} \quad (6.2)$$

The assumption that the prior over model parameters, $\mathcal{P}(\theta)$, is uniform allows the problem to be formulated as a *maximum log likelihood* problem, *ie* one seeks parameters θ^* that maximise $\log \mathcal{P}[\mathcal{D}|\theta]$. The logarithm is included to facilitate the mathematical steps explained below.

Section 6.4 introduces the first of three generative models that have been applied to the problem of finding structure in value functions. Each model incorporates a parametrised structural decomposition of the environment. The model parameters are changed to maximise the probability that the value functions used as data were actually generated by the model.

6.4 The Flat Model

6.4.1 Constructing The Model

There is no unique best probabilistic model. Different models effectively express different prior expectations about how structure might appear. In this and the succeeding sections, a set of very simple probabilistic models is used to extract the relatively simple structure inherent in the navigation tasks described in section 6.2. More sophisticated MDPs may require more sophisticated models.

Value functions $\mathbf{V}^g$ for different goals are treated as being completely independent, so permitting $\log \mathcal{P}[\mathcal{D}|\theta]$ to be written:

$$\log \mathcal{P}[\mathcal{D}|\theta] = \sum_{g} \log \mathcal{P}[\mathbf{V}^g|\theta] \quad (6.3)$$

Each element of $\mathbf{V}^g$, $V^g(s)$, is treated as an independent sample from a simple distribution known as a mixture of Gaussians (*eg* McLachlan & Basford, 1988). This corresponds to the basic intuition about underlying structure – that a particular value is generated by one of a set of value pieces. Hence each data point is considered to have been generated by one of several candidate gaussians:

$$\mathcal{P}(V^g(s)|i,s,g) = \frac{1}{\sqrt{2\pi\omega_i^g}} e^{-(V^g(s)-e_i^g)^2/2\omega_i^g} \tag{6.4}$$

where e_i^g and ω_i^g are the mean and variance associated with gaussian i.

However, *which* gaussian in the mixture is responsible is governed by a weighting $\mathcal{P}(i|s)$ resulting in the following mixture model:

$$\mathcal{P}[V^g(s)|\theta] = \sum_i \mathcal{P}(i|s)\mathcal{P}(V^g(s)|i,s,g) \tag{6.5}$$

Note that these weightings are **not** dependent on the current goal. The different components of the mixture are like spatial *fragments*, and a set, $\mathbf{b}$, of parameters $b_i(s)$ governs the inclusion of state s within fragment i:

$$\mathcal{P}(i|s) = \frac{e^{b_i(s)}}{\sum_j e^{b_j(s)}} \tag{6.6}$$

The probabilistic model of a value function, $\mathcal{P}[\mathbf{V}^g|\mathbf{b}]$, is constructed by optimising e_i^g and ω_i^g for the current goal, g. Hence:

$$\log \mathcal{P}[\mathbf{V}^g|\mathbf{b}] = \max_{\mathbf{e}^g,\boldsymbol{\omega}^g} \left\{ \sum_s \log \left(\sum_i \mathcal{P}(i|s) \frac{1}{\sqrt{2\pi\omega_i^g}} e^{-(V^g(s)-e_i^g)^2/2\omega_i^g} \right) \right\} \tag{6.7}$$

A fragmentation $\mathbf{b}$ is sought that captures the structure of the MDP in a goal independent manner by optimising the resulting expression in equation 6.3 across the whole set of goals that define $\mathcal{D}$. The optimisation is performed by iterating over goals, for each goal performing the following two-stage update:
(1) the appropriately optimised values of $\mathbf{e}^g$ and $\boldsymbol{\omega}^g$ are found for a given goal using a form of EM algorithm (Dempster *et al*, 1977);
(2) $\mathbf{b}$ is then changed using a gradient ascent procedure.

6.4.2 The EM Algorithm

EM is a general, iterative technique for maximum likelihood probability density estimation (Dempster *et al*, 1977; Nowlan, 1990; Specht, 1991). An application of EM begins with

the observation that the optimisation of a likelihood such as $\log \mathcal{P}[\mathbf{V}^g|\mathbf{b}]$ would be simplified if only a set of additional, "hidden" variables were known. For example, in the case of our mixture model (equation 6.5), knowing which gaussian generated which data element would reduce the learning problem to that of independently estimating the mean and variance of a number of gaussians. *Ie*, consider a set of indicator variables, $z_i^{(s)}$, specific to the current goal g (allowing a superscript to be avoided) and equal to 1 if the value of state s was generated by gaussian i, and 0 otherwise. It follows that:

$$\begin{aligned} \mathcal{P}(V^g(s), z_i^{(s)}|\theta) &= \mathcal{P}(i|s)\mathcal{P}(V^g(s)|i,s,g) && (6.8) \\ &= \prod_i \{\mathcal{P}(i|s)\mathcal{P}(V^g(s)|i,s,g)\}^{z_i^{(s)}} && (6.9) \end{aligned}$$

Hence, a quantity can be defined analogous to the quantity to be maximised in equation 6.7:

$$\sum_s \log \mathcal{P}(V^g(s), z_j^{(s)}|\theta) = \sum_s \sum_i z_j^{(s)} \left(\log \mathcal{P}(i|s) + \log \mathcal{P}(V^g(s)|i,s,g)\right) \qquad (6.10)$$

EM suggests finding the expected value of this quantity over the data ($\mathbf{V}^g$), and then maximising the resulting quantity with respect to the parameters ($\mathbf{e}_i^g$ and ω_i^g). Such EM steps are guaranteed not to decrease the likelihood to be maximised (in this case, the expression in curly brackets in equation 6.7).

For the **E-Step**, note first that:

$$\begin{aligned} E\left[z_i^{(s)}\right] &= \mathcal{P}(z_i^{(s)} = 1|s, V^g(s), \theta) && (6.11) \\ &= \frac{\mathcal{P}(V^g(s)|z_i^{(s)} = 1, s)\mathcal{P}(z_i^{(s)} = 1|s)}{\mathcal{P}(V^g(s)|s)} && (6.12) \\ &= \frac{\mathcal{P}(V^g(s)|i,s)\mathcal{P}(i|s)}{\sum_j \mathcal{P}(V^g(s)|j,s)\mathcal{P}(j|s)} = h_i^{(s)} && (6.13) \end{aligned}$$

Equation 6.11 follows from the fact that $z_i^{(s)}$ is an indicator variable. Equation 6.12 follows from Bayes' rule. Hence the value of the expectation is usually referred to as a *posterior probability*, and denoted $h_i^{(s)}$. The θ notation has been dropped for convenience from equations 6.12 and 6.13. The above expectation allows us to compute the expected value of the quantity in equation 6.10, which is commonly referred to as the Q function:

$$E\left[\sum_s \log \mathcal{P}(V^g(s), z_j^{(s)}|\theta)\right] = \sum_s \sum_i h_i^{(s)} \left(\log \mathcal{P}(i|s) + \log \mathcal{P}(V^g(s)|i,s,g)\right)$$
$$= Q \tag{6.14}$$

The **M-Step** involves finding new parameters to maximise this quantity, *ie* finding, for the $n+1$'th iteration:

$$\{e_i^g, \omega_i^g\}^{(n+1)} := \arg\max_{\{e_i^g, \omega_i^g\}} \left\{ \sum_s h_i^{(s)} \log \mathcal{P}(V^g(s)|i,s,g) \right\} \tag{6.15}$$

where the posterior probability $h_i^{(s)}$ is calculated using the parameter values from the previous iteration, *ie* $\{e_i^g, \omega_i^g\}^{(n)}$. Note the problems for individual gaussians have now been separated. The following rules follow straightforwardly by taking derivatives with respect to the parameters, and setting to zero. For the mean parameter e_i^g:

$$\{e_i^g\}^{(n+1)} := \frac{\sum_s h_i^{(s)} V^g(s)}{\sum_s h_i^{(s)}} \tag{6.16}$$

and for the variance parameter ω_i^g:

$$\{\omega_i^g\}^{(n+1)} := \frac{\sum_s h_i^{(s)} (V^g(s) - \{e^g\}_i^{(n+1)})^2}{\sum_s h_i^{(s)}} \tag{6.17}$$

E and M phases are alternated for about 10 iterations, which is usually ample to get adequately near convergence. The resulting, optimised parameter values will be referred to as $\overset{*}{e}{}_i^g$ and $\overset{*}{\omega}{}_i^g$. A deficiency in the EM algorithm is that variances can decrease to zero if allowed to – an EM version of data overfitting. It is usually necessary to prevent variances from getting too small, and so decreasing changes in any ω_i^g were capped at 0.5. This can be interpreted as a form of regularisation. A second deficiency in common with many optimisation methods is that while a local maximum is assured, a global one is not. At present, the only way in which this problem is addressed is by keeping the variance parameters ω_i^g large at the beginning of each set of iterations. Other techniques might be to use more sophisticated initialisation methods or multiple restarts to avoid bad local maxima.

6.4.3 Gradient Ascent

Having found $\overset{*g}{e}_i$ and $\overset{*g}{\omega}_i$ for a particular goal, the next step is to alter the fragmentation parameters $b_i(s)$. First, the following quantities can be defined, to make life easier:

$$p_i^g(s) = \frac{1}{\sqrt{2\pi\overset{*g}{\omega}_i}} e^{-(V^g(s)-\overset{*g}{e}_i)^2/2\overset{*g}{\omega}_i} \tag{6.18}$$

$$\psi_i(s) = \mathcal{P}(i|s) \tag{6.19}$$

$$\mathrm{P}^g(s) = \sum_i \psi_i(s) p_i^g(s) \tag{6.20}$$

Then, a gradient ascent procedure can be specified, starting from the global likelihood function (equation 6.3) and that for a single value function (equation 6.7):

$$\frac{\partial}{\partial b_k(s)} \log \mathcal{P}[\mathcal{D}|\theta] = \frac{\partial}{\partial b_k(s)} \sum_g \sum_{s'} \log \mathrm{P}^g(s') \tag{6.21}$$

$$= \sum_g \frac{\partial \log \mathrm{P}^g(s)}{\partial b_k(s)} \tag{6.22}$$

Dropping the dependence on s for the three simplifying quantities defined above, this is equal to:

$$= \sum_g \sum_i \frac{\partial \log \mathrm{P}^g}{\partial \psi_i} \frac{\partial \psi_i}{\partial b_k(s)} \tag{6.23}$$

$$= \sum_g \sum_i \frac{p_i^g}{\mathrm{P}^g} \frac{\partial \psi_i}{\partial b_k(s)} \tag{6.24}$$

Because all ψ_i have a dependence on b_k, ψ_k and $\psi_{i\neq k}$ are considered separately as follows:

$$\frac{\partial}{\partial b_k(s)} \log \mathcal{P}[\mathcal{D}|\theta] = \sum_g \left(\frac{p_k^g}{\mathrm{P}^g} \psi_k (1-\psi_k) - \sum_{i\neq k} \frac{p_i^g}{\mathrm{P}^g} \psi_i \psi_k \right)$$

$$= \sum_g \left(\frac{p_k^g \psi_k}{\mathrm{P}^g} - \psi_k \right) \tag{6.25}$$

$$= \sum_g \left(\overset{*(s)(g)}{h}_i - \psi_k \right) \tag{6.26}$$

Note that now the posterior probability's dependence on g is made explicit with a slight change of notation. Thus, changes in $b_k(s)$ proportional to the right hand side of equation 6.26 act to minimise the difference between the prior and posterior probabilities that

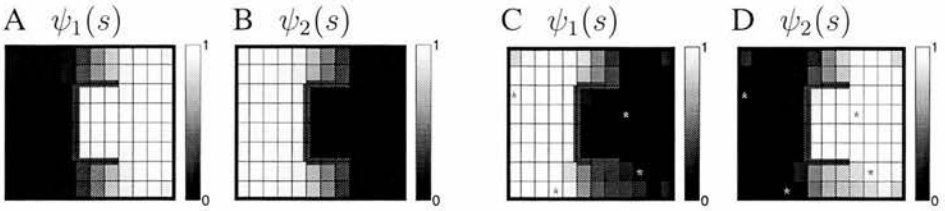


Figure 6.4: Structural decompositions into two flat regions of the maze in figure 6.2A. A;B) $\psi_1(s)$ and $\psi_2(s)$ when trained on all goals. C;D) $\psi_1(s)$ and $\psi_2(s)$ when trained on the four goals shown by the stars. The internal barriers are shown in gray to distinguish them from $\psi_i(s)$.

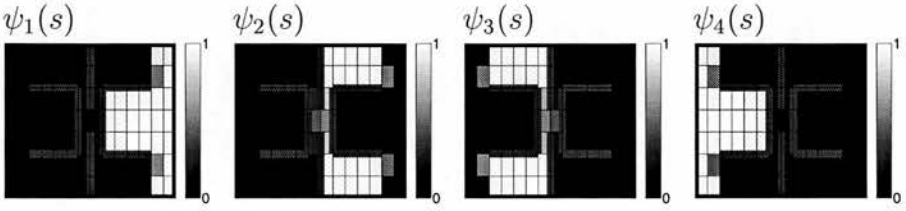


Figure 6.5: Structural decompositions into four flat regions ($\psi_1(s), \psi_2(s), \psi_3(s)$ and $\psi_4(s)$) of the maze in figure 6.2B.

the values for state s (across several goals) were generated by the k'th fragment (which was allowed to optimise its parameters for each goal).

To recapitulate, for each goal the fragmentation parameters $\mathbf{b}$ are changed according to

$$\Delta b_i(s) \propto \overset{*(s)(g)}{h_i} - \psi_i(s) \tag{6.27}$$

where $\overset{*(s)(g)}{h_i}$ is defined according to equation 6.13 using the values $\overset{*g}{\mathbf{e}}$ and $\overset{*g}{\boldsymbol{\omega}}$ produced at the end of the EM procedure. The whole process is repeated for either randomly or systematically chosen goal positions.

6.4.4 Results For The Flat Model

Training On All Goals

Figure 6.4A;B shows the results of applying the flat model to the maze of figure 6.2A in the case that there are two fragments. The figures show $\psi_1(s)$ (A) and $\psi_2(s)$ (B). The unsupervised learning procedure has taken the states and found a structural decomposition that seems appropriately sensitive to the structure of the environment, taking account of spatial proximity and, where appropriate, the barriers.

Figure 6.5 shows the decomposition that results from applying the flat model to the two-room maze in figure 6.2B. Once again, the decomposition is appropriately sensitive to the underlying structure of the task.

Training On A Subset Of Goals

The ability to capture underlying structure on the basis of far fewer goals than all possible is clearly fundamental to the problem of goal generalisation. To confirm that some aspects of the structure of the environment are apparent given training on only a few goals, figure 6.4C;D show $\psi_1(s)$ and $\psi_2(s)$ in the case that the optimal value functions for just four goals (the grey stars) were used. The resulting structural decomposition is slightly different from the one that develops based on all the goals, and indeed would be more different if the four goals had been confined to just one of the 'regions' defined in figure 6.4A;B. Nevertheless, the basic structure of the environment remains evident.

6.5 The Linear Model

The flat model employs very little prior information in order to find good decompositions. Most applications of unsupervised learning involve the use of more or less sophisticated prior information to shape the representations. An obvious prior (apparent, for instance, in the value functions of figure 6.3) is that there should be contiguous regions of the environment in which the value function is essentially linear in the location $\mathbf{u}(s)$ of state s (where $\mathbf{u}(s)$ is a two dimensional vector corresponding to the two spatial dimensions of the maze). Moreover this is compatible with the coordinate information which can easily be learned in an environment (see Chapter 4), though there are likely to be restrictions on when coordinates are available, such as only within environmental structures of a certain size, *ie* a global coordinate system cannot be assumed.

In the linear model, the flat Gaussian model for fragment i of equation 6.4 is replaced by:

$$\mathcal{P}(V^g(s)|i,s,g) = \frac{1}{\sqrt{2\pi\omega_i^g}} e^{-(V^g(s)-\mathbf{d}_i^{g\,T}\mathbf{u}(s)-e_i^g)^2/2\omega_i^g} \tag{6.28}$$

with a set of extra parameters $\mathbf{d}_i^g$ governing the slope of the linear dependence.

Because there are now three parameters governing the mean of each gaussian, optimising all three together requires a slightly more complex procedure than that of equation 6.16. First two vectors are defined, a parameter vector, $\mathbf{M}_i^g$, and a corresponding state vector, $\mathbf{S}(s)$:

$$\mathbf{M}_i^g = \begin{bmatrix} d_i^g{}_1 & d_i^g{}_2 & e_i^g \end{bmatrix}^T \tag{6.29}$$

$$\mathbf{S}(s) = \begin{bmatrix} \mathbf{u}_1(s) & \mathbf{u}_2(s) & 1 \end{bmatrix}^T \tag{6.30}$$

The remaining step is to solve $\nabla_{\mathbf{M}_i^g} \mathcal{P}(V^g(s)|i,s,g) = 0$, where the mean of this distribution is now expressed as $\mathbf{M}_i^g \cdot \mathbf{S}(s)$.

The update rule of equation 6.16 therefore becomes:

$$\{\mathbf{M}_i^g\}^{(n+1)} := \left(\sum_s h_i^{(s)(g)} \mathbf{S}(s)\mathbf{S}^T(s)\right)^{-1} \left(\sum_s h_i^{(s)(g)} V^g(s)\mathbf{S}(s)\right) \quad (6.31)$$

where $(.)^{-1}$ indicates the matrix inverse. The update rule for the variance parameters becomes:

$$\{\omega_i^g\}^{(n+1)} := \frac{\sum_s h_i^{(s)} \left(V^g(s) - \{\mathbf{M}_i^g\}^{(n+1)} \cdot \mathbf{S}(s)\right)^2}{\sum_s h_i^{(s)}} \quad (6.32)$$

6.5.1 Results For The Linear Model

Difference Between Linear and Flat

Figure 6.6 shows the result of fitting five linear fragments to the maze of figure 6.2B. The principal difference between this and the flat decomposition is that the linear pieces clearly favour separating the area within each barrier from other areas. This follows naturally from a fitting scheme which can take advantage of the slopes of the value function for areas within and outside of a barrier, slopes which are often different. By contrast, the flat fragments show how absolute values are more likely to cluster into regions on one or the other side of the barrier.

Figure 6.6 also makes the point that only four of the five fragments are substantially employed – the mirror image of $\psi_1(s)$ is almost the combination of $\psi_3(s)$ and $\psi_4(s)$.

Scaling of Fragmentation

Further insight into the nature of these decompositions can be gained by fitting the same number of fragments to a similarly structured environment that contains four times as many states (392 states). The resulting decomposition is essentially equivalent to that for the smaller state space (figure 6.7), demonstrating that the decompositions scale, in a sense, with the essential complexity of the environmental structure rather than with the size of the state space.

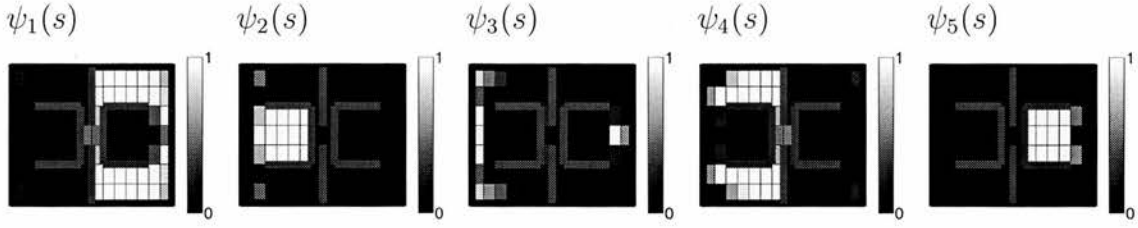


Figure 6.6: Structural decompositions into five linear regions ($\psi_1(s), \psi_2(s), \psi_3(s), \psi_4(s)$ and $\psi_5(s)$) of the maze in figure 6.2B.

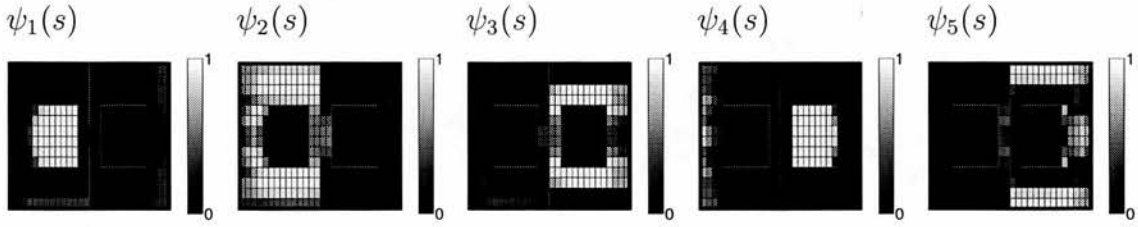


Figure 6.7: Structural decompositions into five linear regions ($\psi_1(s), \psi_2(s), \psi_3(s), \psi_4(s)$ and $\psi_5(s)$) of a maze structurally similar to figure 6.2B, but made up of 392 states.

6.6 The Hierarchical Model

The third and final model is an attempt to capture hierarchical structure in environments. The mixture models considered so far put pressure on the resulting decompositions to be *exclusive*, and this is demonstrated in the figures, for which the mixture components $\psi_i(s)$ are close to 0 or 1 for most states s. For a more hierarchical fragmentation, decompositions are sought that are exclusive in terms of fragments of the same level, but where states are free to belong simultaneously to many fragments at different levels, reflecting for example their position in rooms, in particular parts of rooms, and the like.

There are various ways to place states simultaneously in multiple fragments. By way of introduction, the following simple scheme is first investigated, which uses a form of pyramidal representation first suggested in the context of computer vision (*eg* Burt & Adelson, 1983). Having decomposed the value functions as best as possible at one level, a fragmentation at a lower and potentially more detailed level is generated by fitting the residual error from the first fragmentation. The fitting procedure is non-linear, involving iterations of the EM algorithm, and so the multiple fragmentations are not the same as the trivial case of fitting more fragments in the first place. Further, there is no reason for the fragments at one level to be strictly included in the fragments at higher levels, although this can emerge for certain environments. This scheme will be referred to as a *loose* hierarchical decomposition.

6.6.1 The Loose Hierarchical Model

For convenience, the problem of learning a two-level hierarchy is considered. Furthermore, only flat fragments in each level are considered, although the extension to the linear case for the lower level is straightforward. The approach is to amend the procedure described at the end of section 6.4.1, so that for each goal the following four-stage update is performed:
(1) the optimised parameters for the *top* level, $\overset{*g}{e_i}$ and $\overset{*g}{\omega_i}$, are found using EM;
(2) the *top* fragment parameters, $b_i(s)$, are changed using gradient ascent;
(3) optimised parameters for the *bottom* level are found, using EM and the new data set of residual value functions, $\mathbf{U} = \left(\mathbf{V}^g - \overset{*g}{\mathbf{V}}\right)$ where

$$\overset{*g}{V}(s;\boldsymbol{\psi},\overset{*g}{\mathbf{e}}) \;=\; \sum_i \psi_i(s)\overset{*g}{e_i} \tag{6.33}$$

(4) fragmentation parameters for the *bottom* level are changed using gradient ascent with the new dataset.

The scheme as described is not quite correct, because the top-level fragmentation is used to prescribe the residual data set *before* it has converged to its optimum. However, this computational convenience is achievable because, having potentially fewer parameters to optimise, the top-level fragmentation can be found relatively quickly, and further this fragmentation remains independent of that of the lower level. A second fault of the scheme is that the residual variances, $\overset{*g}{\boldsymbol{\omega}}$, are ignored. These may include information about the quality of fit at the top-level.

6.6.2 Results For The Loose Hierarchical Model

Figure 6.8 shows the consequence of fitting flat fragments (*ie* without linear 'pieces') for the simple maze in figures 6.2A;D in the case that the top level flat fragments are those shown in figure 6.4A;B. Figure 6.8A;B show that the appropriate second-level fragmentation is roughly orthogonal to the first level fragments, which is a natural consequence of the sort of residual errors left by the first level flat fragments. Even the details of the second level fragmentation (for instance, the behavior near to the corners of the maze) turn out to be appropriate. Figures 6.4C;D;E show the additive construction of the final model of the value function (shown in figure 6.3A) for a particular goal. Figure 6.4C shows the component from the upper level fragments; figure 6.4D the component from the lower level fragments; and figure 6.4E shows their sum. One can see from figure 6.4E that there is substantial residual error in the details of the value function – naturally, since the representation has been reduced from 96 numbers to just 8 (including the variance terms).

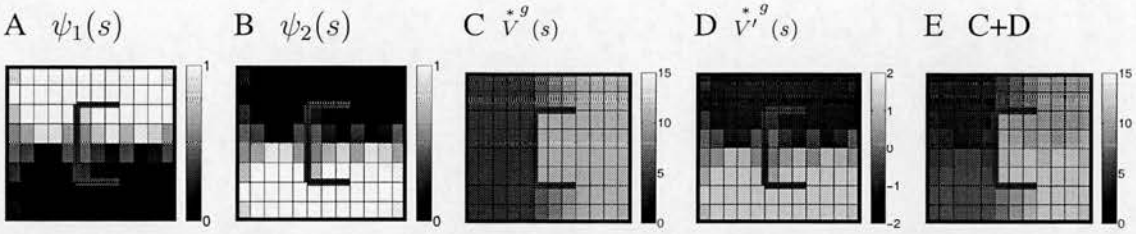


Figure 6.8: Second level model for the maze in figure 6.2A – the first level fragmentation is shown in figure 6.4A;B. A;B) Fragmentation $\psi_1(s)$ and $\psi_2(s)$. C) Contribution to modeling the optimal value function when the goal is at location $(3, 3)$ (shown in figure 6.3A) from the first level fragments of figure 6.4A;B. D) Contribution to modeling the value function from the second level fragments of (A;B) in this figure. E) Sum total model of the value function in figure 6.3A.

6.6.3 The Cooperative Hierarchical Model

A more complicated environment suggests a more appropriate hierarchical decomposition. In many cases, it is likely to be useful to have a stricter hierarchical decomposition, in which fragments at a lower level are strictly included within fragments at higher levels. For example, the maze of figure 6.2C suggests a strict decomposition into rooms and within rooms into fragments. Moreover, the requirement of learning to place 8 fragments as is made by the loose hierarchical model, even given that the residual values are being used rather than the full values, is likely to be made much easier with hierarchical information specifying, for example, that pairs of fragments belong to a certain room.

One method of fitting fragments hierarchically is the *hierarchical mixture of experts* model of Jordan and Jacobs (1993), in which the choice of expert (*ie* the question in the present case of which gaussian is responsible for a given data point) is modeled as a hierarchical decision tree. The problem with this approach is that only the mixture part of the model is actually hierarchical – the experts themselves reside at the leaves of the decision tree. Here, by contrast, a *cooperative* hierarchy of experts is desired, in which the value of a given state is the additive result of several experts at different resolutions.

The loose hierarchical model provides the inspiration for a second model, which will be referred to as the *cooperative hierarchical model*. First, the value functions are decomposed as well as possible at the top level of the hierarchy, producing a residual error function. This is exactly the same step as for the loose hierarchical model. Subsequently however, instead of fitting a number of normal fragments to this residual error, a hierarchical mixture of experts model is used to fit the lower level fragments to the residual error. Greater prior knowledge is assumed, in the form of a fixed number of low-level fragments per high-level fragment. The desired (and found) result is that within a high-level fragment, low-level fragments are assigned as if the high-level fragment was all the environment that existed.

The probabilistic model of the residual value function, $\mathbf{U}^g$, is now:

$$\log \mathcal{P}[\mathbf{U}^g|\mathbf{b}] = \mathcal{K} + \max_{\mathbf{e}^g,\boldsymbol{\omega}^g} \left\{ \sum_s \log \left(\sum_i \mathcal{P}(i|s) \sum_{j \in \mathcal{F}_i} \mathcal{P}(j|i,s) \frac{1}{\sqrt{2\pi\omega_{ij}^g}} e^{-(U^g(s)-e_{ij}^g)^2/2\omega_{ij}^g} \right) \right\}$$

where $\mathcal{F}_i$ refers to the (pre-ordained) set of lower level fragments permitted to reside within the i'th high-level fragment, and $\psi_{j|i}(s) = \mathcal{P}(j|i,s)$ is the fragment membership probability (parametrised by b_{ij}) for the j'th lower-level fragment which is conditional on the higher-level fragment, and $\psi_i(s) = \mathcal{P}(i|s)$ is the fragment membership probability (parametrised by b_i) for the higher-level fragment.

EM can be used to perform the maximisation, as before. The E-step defines a Q function, analogous to equation 6.14:

$$Q = \sum_s \sum_i \sum_j h_{ij}^{(s)} \left(\log \psi_i(s) + \log \psi_{j|i}(s) + log\mathcal{P}(U^g(s)|i,j) \right) \tag{6.34}$$

where:

$$h_{ij}^{(s)} = \frac{\psi_i(s)\psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)}{\sum_i \psi_i(s) \sum_j \psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)} \tag{6.35}$$

Two related posterior probabilities can also be defined:

$$h_i^{(s)} = \frac{\psi_i(s) \sum_j \psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)}{\sum_i \psi_i(s) \sum_j \psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)} \tag{6.36}$$

$$h_{j|i}^{(s)} = \frac{\psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)}{\sum_j \psi_{j|i}(s)\mathcal{P}(U^g(s)|i,j)} \tag{6.37}$$

Note that $h_{ij}^{(s)} = h_i^{(s)} h_{j|i}^{(s)}$.

The M-step is used exactly as before, to optimise only the "expert" parameters, in this case e_{ij}^g and ω_{ij}^g, leading to the following updates, analogous to equations 6.16 and 6.17:

$$\{e_{ij}^g\}^{(n+1)} := \frac{\sum_s h_{ij}^{(s)(g)} U^g(s)}{\sum_s h_{ij}^{(s)(g)}} \tag{6.38}$$

$$\{\omega_{ij}^g\}^{(n+1)} := \frac{\sum_s h_{ij}^{(s)(g)} (U^g(s) - \{e_{ij}^g\}^{(n+1)})^2}{\sum_s h_{ij}^{(s)(g)}} \tag{6.39}$$

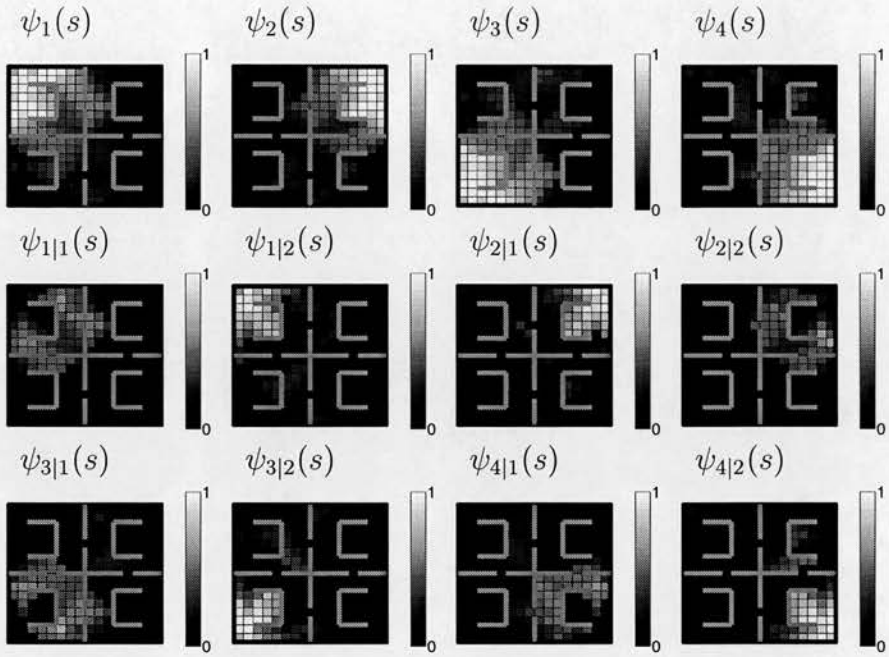


Figure 6.9: Cooperative hierarchical structural decomposition for the four room environment: the top row shows the four fragments in the upper level of the hierarchy, and the bottom two rows show the eight fragments in the lower level of the hierarchy, trained on the residual errors of the upper level, using a hierarchical mixture model.

A gradient ascent procedure analogous to that of equation 6.26 can be determined, but importantly weight changes are weighted by the (maximised) posterior probability of the high-level fragment, $\overset{*(s)(g)}{h_i}$, as follows:

$$\Delta b_{ij}(s) \propto \overset{*(s)(g)}{h_i} \left(\overset{*(s)(g)}{h_{j|i}} - \psi_{j|i}(s) \right) \tag{6.40}$$

Note that changes are not now made to high-level fragment parameters, $b_i(s)$.

6.6.4 Results For The Cooperative Hierarchical Model

Figure 6.9 shows the consequence of fitting a hierarchy of four flat fragments in the upper level, and eight flat fragments at the lower level using the cooperative hierarchical model. Note that one's natural intuition that the correct higher level decomposition should find the four rooms is correct. Further, the hierarchical mixture component makes the problem of finding lower-level fragments fairly easy, since learning is almost completely restricted by the higher level fragmentation to, in each case, a single room.

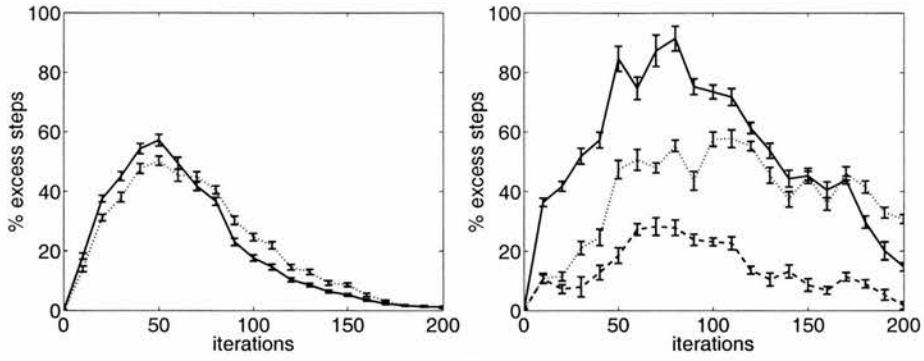


Figure 6.10: Faster actor-critic learning with the augmented representation. (A) A single-level, linear decomposition (5 fragments) was used in the 2-room environment of figure 6.2B. Mean numbers of steps were generated every ten trials by averaging over a number of non-learning trials (one trial per possible start state). The process was repeated for every possible goal, and the mean taken, and this constituted a single "run". The figure is then the product of 30 such runs. The solid line shows the *difference* between the performance of the conventional actor-critic, and the augmented actor-critic, as a percentage of the performance of the conventional actor-critic (percent extra steps). Likewise, the dotted line shows the percent extra steps for an actor-critic adorned with a randomly assigned fragmentation. (B) The cooperative hierarchical decomposition of figure 6.9 was tested in the 4-room environment of figure 6.2C. The curves (for 5 runs) were generated in a similar manner to those of figure A, and show the percent extra steps taken by a standard actor-critic (solid line), an actor-critic augmented with a randomly assigned, two-level fragmentation (dotted line), and an actor-critic augmented with just the single level of fragmentation (8 fragments only; dashed line) from figure 6.9.

6.7 The Actor-Critic

The unsupervised learning techniques are intended to find decompositions which reflect the underlying structure of environments. However, it is not necessarily the case that the decompositions that are found will be of any use for reinforcement learning. In particular, unsupervised learning has been applied to *optimal* value functions, whereas to solve MDPs, one is for the most part considering *sub-optimal* value functions associated with current (but changing) predictions. Since, however, the decompositions found appear to reflect underlying structure, it might be hoped that they would be of use in a learning context as well.

To address this issue, multiple-goals MDPs were addressed using temporal difference (TD) learning within an actor-critic architecture. For two environments (the 2-room of figure 6.2B and the 4-room of figure 6.2C), an actor-critic was trained separately on each of the MDPs defined by all possible goals (*ie* a goal for every state), using either a conventional state representation, or the *augmented* representation resulting from having both

conventional states and fragments as concurrent inputs to the system. As a control condition, randomly assigned fragmentations were also used.

In effect, fragments and rooms were used to define a distributed representation. For example, for the hierarchical decomposition of figure 6.9, the estimated value $U(s)$ of a state s is given by:

$$U(s) = w_s + \sum_f \psi_f^{\text{FRAG}}(s) w_f^{\text{FRAG}} + \sum_r \psi_r^{\text{ROOM}}(s) w_r^{\text{ROOM}} \tag{6.41}$$

where w_s is the critic weight for state s, w_f^{FRAG} is the critic weight for fragment f, and w_r^{ROOM} is the critic weight for room r. Note that the fragments are treated completely separately in this simple simulation, *ie* the "ownership" of fragments by rooms is not incorporated into the learning. A similar calculation specifies action preferences in the actor (with three different sets of weights). Finally, in applying learning equations 3.9 and 3.18, the change to fragment and room weights is proportional to the membership probability in each case (*ie* $\psi_f^{\text{FRAG}}(s)$ or $\psi_r^{\text{ROOM}}(s)$).

The learning rate constants were optimised to the following values. For weights in the linear decomposition simulation: state (critic; actor) = (.5, 2), fragment (critic e_f; critic slope $\mathbf{d}_f$; actor)=(.03, .00015, .25) for the proper fragmentation, and (.03, .00015, .12) for the random fragmentation. For weights in the cooperative hierarchical decomposition study: state = (.375, 1.5) in all cases except for unaugmented learning without any fragment information, which was (.5, 2); fragment = (.034, .13); room = (.007, .028).

For each case considered, the percent average *extra* steps to the goal, over the learner with the augmented representations, was measured. The use of the decompositional information significantly speeded up learning in both the 2-room (figure 6.10A) and 4-room (figure 6.10B) environments. Moreover, the appropriateness of the additional information is paramount, as demonstrated by the poor learning with a randomly assigned decomposition. In both cases, these speed-ups are based on just the relatively tiny amount of extra information provided by the fragmentation.

6.8 Discussion

A new method has been presented for extracting structure in MDPs in which unsupervised learning techniques are applied to a collection of value functions, decomposing them into their elemental pieces. Fragmentations inferred on the basis of only a handful of goals have been shown to capture underlying structure. Complete fragmentations significantly speed up simple temporal difference learning methods for finding optimal policies.

There are important limitations with the method as it has been investigated so far. The first is that the methods stop short of a full solution to the problem of multiple-goals MDPs – because they do not capture the fact that as well as values being correlated for nearby states across many goals, they are also likely to be correlated for nearby goals. This was, in fact, stated in section 6.1.1 as the hierarchical principle. Not only is this information not utilised in learning the clustering, but it is ignored in the simple actor-critic learning which was investigated. A more general form of actor-critic would incorporate a multiplicative representation of both current location and goal, in which *both* variables were represented using the multi-resolution representation afforded by the discovered structure. This remains for future work.

The second limitation is that values in themselves are, as noted in the introductory section 6.1.3, *not* necessarily the only suitable source of clustering information. It is certainly not possible to hypothesise a neural requirement for value information with which to form clustered representations of the sort found. For example, a clustering algorithm might conceivably be based on reports of success or failure from a coordinate-based controller – thus achieving clustering quite independently of any goal- or reward-related information. A more subtle point is that the Gaussian models imply a cost for incorrect values that varies with the square of the residual error. In fact, the exact quantities of the value functions do not matter, it is really only important that the resulting policies are correct. A more sophisticated model might somehow take the quality of the resulting policy into account.

The actor-critic has played a dual role in the present study by confirming intuitions about the clustering, and exploring an attractively simple way in which it might be used. First, although the fragmentations reported appear to capture structure, as far as can be discerned by looking at the results, and certainly also reduce the error associated with approximating the value functions, nevertheless without the actor-critic there would be no direct proof that the fragmentations were in any sense correct. In fact, actor-critic learning using the optimised fragmentations produced credible improvements over that using random fragmentations, *ie* using a similar number of learning parameters. This is particularly important to establish, because although the fragmentations were determined from optimal value functions, during learning they must be used to support sub-optimal values as well. The underlying belief is that these sub-optimal values are just as dependent on environmental structure as the optimal values, and so learning should be aided. Second, the actor-critic has been proposed as a simpler and more neurally plausible alternative to explicitly hierarchical methods such as feudal learning. Uses of structural information within a simple RL context have been proposed, but have sought to aggregate states, that is, in the terms that have been used here, *replace* low-level states with aggregated ones. The dangers of doing this, in terms of inducing partial observability, are well understood (Singh *et al*, 1994). There is however

little theory as to how using structure to *augment* a representation of state can help, or hinder, learning. While this approach does nothing to reduce the number of states involved in planning (indeed this is increased), nevertheless augmentation appears to be a safer and more flexible strategy.

6.8.1 Related Work in RL (I): From Temporal To Structural Abstraction

Much recent work has focused on similar issues to those addressed in the present work: the problem of large state spaces, and generalisation between similar MDPs differing only, for example, by which state is the goal state. Two broad classes of approach can be discerned. *Temporal abstraction* starts from the idea that it may be inefficient always to plan actions at the finest timescale, and so considers, for example, how temporally extended sequences of actions may be evaluated, and how useful forms of such sequences may be determined. *Structural abstraction* starts from a similar idea to the hierarchical principle of section 6.1.1 (and not necessarily the structural principle), that often across a range of problems subsets of states will have similar values or require similar actions. For navigation, this corresponds to the idea that it may be inefficient always to plan at the smallest spatial scale.

Temporal abstraction is the safest place to start. An extension to the theory of MDPs allows for transitions of variable duration, and is often referred to as the theory of semi-MDPs, or SMDPs. In particular, RL algorithms such as TD learning work in SMDPs; for TD the only differences are that the single return received after a transition is replaced by the total reward received during the duration of the transition, and the discount factor is now a function of the duration (*ie* of the time-length of the duration). This theory allows for the following observation: under some policy π, predictions made about a temporally extended transition, *eg* the probability of a transition from some state A to some other state B not immediately attainable from it, can be made either in terms of the low-level state transitions, or in terms of a high level SMDP transition between A and B for which the intermediate states are invisible – it does not matter because the predictions will be the same (figure 6.11).

This theory has been developed to allow for the specification of temporally-extended actions so as to speed up learning. In one example, Singh (1992b) demonstrated that learning a set of "abstract actions", effectively separate policies, for getting to 3 goal states, subsequently speeded up learning a composite task made up of successive visits to the same goal states. In a second example, Precup and Sutton (1998) consider a simple grid-world composed of four rooms with doorways between them. For a set of "options" which took an agent from a state within a room to one of the two doorways available from that room, again effectively policies but here with a restricted domain of states, Precup and Sutton

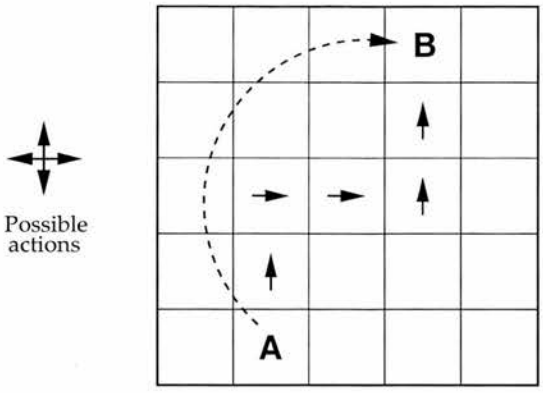


Figure 6.11: An extended action can be expressed as a set of single-step actions, or as a single transition from start state to end state.

demonstrate speeded up learning over that using just lowest level actions alone.

These examples make it clear that the specific choice of temporally-extended actions is key to their success, and that these methods are far from providing a comprehensive treatment of how appropriate actions might be learned. The question remains, however, if the temporally-extended actions lead to the same evaluations as the low-level actions, why is it quicker using the former? The answer is that temporally-extended actions effectively work by focusing an agent's exploration on certain key objectives.

The idea that *a priori* knowledge about the task might be used to effectively constrain exploration has been taken much further using Hierarchies of Abstract Machines (HAMs; Parr and Russell, 1998). A HAM is a finite-state machine operating alongside the MDP of interest, and consisting of a set of machine states, which can be one of four types: action states specifying an action in the MDP; call states which call other machines (allowing hierarchical structure); stop states which cease the machine; and choice states which non-deterministically select a new state. On a theoretical level, it is possible to define a new MDP whose state space is the cross-product of the machine with the original MDP, and provided certain conditions are met with regard to how transitions are handled concurrently in the two subspaces, optimal policies in the larger MDP correspond to optimal policies in the original. However, most of the states in the larger MDP can be ignored, since either they are unattainable, or they exist as transient states between choice points, and can be effectively got rid of (using the SMDP theory). Nevertheless, one can imagine that every choice machine state would need to be tried in every original MDP state – there is no pre-specified way of generalising across states. The reason why HAMs can be efficient is, again, because they constrain the exploration of the agent. In particular, the strict manner in which control is passed from machine to machine implies that certain action complexes will be tried until they are finished, *ie* reach some stopping criteria, and this alone is a useful constraint on the actions of the agent. HAMs are a very flexible tool for incorporating knowledge about

the task. However, as Parr and Russell note, even HAMs would be enhanced if structural knowledge could be incorporated – otherwise many temporally extended actions will be attempted from states where they are inappropriate.

The notion that structural abstraction should partner temporal abstraction has been used by Hauskrecht *et al* (1998). A formal decompositional technique for MDPs is used (Dean and Lin, 1995) to specify regions and sub-goal states. Within each region, policies can be defined in terms of moving to the sub-goal states. These act as temporally abstract actions to speed up learning policies when learning several similar MDPs in which only the goal position moves. However, the decompositional technique relies on a rather rigid specification of sub-goals and regions; the convenient examples they consider appear to have one subgoal allotted to each region. In general, however, environments are likely to be more complex than this. Theirs is not a sufficiently general form of structural learning.

A more direct way still of using structural abstraction in the specification of policies has been investigated by Thrun and Schwarz (1995). Their technique is very similar to that pursued in this chapter, with the added felicity that they are working directly with policies, while we use the surrogate of value functions. They consider a set of similar problems, in fact navigational MDPs which differ only in goal position, and attempt to learn re-usable policy pieces for subsets of tasks. Given a set of optimal Q-values for all the MDPs (instead of the optimal values which we used), they choose parameters to minimise a certain cost function. The parameters specify re-usable policies, states in which they act, and tasks for which they are suitable. The cost function is a combination of *description length* in which individually specifying policies for every task is penalised, and *performance loss* in which re-using policies inappropriately is penalised. However, the combination is somewhat *ad hoc* – it is not clear why the two quantities can be compared in this way, within a single metric. Nevertheless, this is an interesting approach to learning structure, and provides a counter-point to that presented in this chapter.

Finally, an alternative to hierarchical abstraction which explicitly attempts to capture underlying structure is the Parti-Game algorithm of Moore and Atkeson (1995). They consider starting with no state structure at all, *ie* a continuous N-dimensional space, and constructing a state space with as few states as possible. Parti-game only suggests a solution to this problem, since the algorithm is model-based (*ie* uses DP methods explicitly) and deterministic, in that it calculates values for states on the basis of a continually updated database of experienced state transitions. Further, task-specific knowledge is assumed of the direction of the goal (such as a coordinate system might provide). The difficulties arise when starting at a particular location, setting off at the appropriate bearing to get to the goal and instead meeting a barrier. Parti-game starts with a single, environment-wide state, and a goal state consisting of a predefined goal region. A database is continually updated with entries of

the form (i, j, k) indicating that a transition was attempted from state i to state j, and that the resulting state was k. A simple example of the effect of barriers on these entries is that in each case, k can be equal to either of i or j. Values are then assigned to states by a minimax algorithm which maximises the value of states subject to the constraint that, given any choice of next state, the worst possible outcome results, as indicated by the transition database. This is equivalent to one attempting to maximise the value of states given that a malicious opponent is allowed to place one in the worst possible position within one's current state. Clearly, for many states the resulting value will be infinitely low (facilitated by the assumption of fixed negative returns per transition, an absorbing goal state, and initial value estimates of zero). This is interpreted by Parti-Game as a need for better resolution. The set of states is determined that are themselves infinitely low in value but also *border* states that are not. Both these states and the states they border are each then bisected along their longest axis, creating two states. Thus the algorithm tends to increase resolution close to the edges of barriers, or close to apertures within barrier-like structures, and on both sides of the barriers. Areas which do not have barriers tend to remain at a lower resolution. The entire process of storing transitions, determining values and changing the state space is repeated at each step of a path to the goal.

Moore and Atkeson demonstrate that Parti-Game learns faster than more conventional algorithms (including Q-learning) on a two dimensional navigation problem. However, the true advantage of the approach is clearer with higher dimensional spaces. In a nice example, they consider the problem of manoeuvering a snake-like arm of fixed length but composed of N segments (with $N - 1$ joints) around an obstacle and into a "goal" position. Different problems can be defined by different values of N. The number of states which Parti-Game creates to solve these problems increases apparently somewhat less than linearly with N. By contrast, conventional DP-based methods would require that the number of states increase exponentially with N, with learning times rising accordingly. Parti-Game can apparently extract the underlying structure of the problem without succumbing to the curse of dimensionality, but also without having to use a hierarchical description of state.

6.8.2 Related Work in RL (II): Explicitly Hierarchical Control

The following discussed methods are not particularly plausible from a neural perspective, but are interesting for the forms of hierarchical structure which they employ. In a sense, these are the natural rivals for our methods. Feudal learning is one such explicit hierarchy, and has been discussed at length in section 6.1.2.

Dayan (pers. comm.) noted an inefficiency in FL which has been addressed in an extended scheme called the MAXQ architecture (Dietterich, 1998). At the end of learning, a rela-

tionship can be expected between the Q-values learned by a controller and the Q-values learned by that controller's vassals, but feudal learning is oblivious to this, with apparently separate MDPs at each level. Furthermore, this relationship can be expected to be similar to that between the evaluation of temporally extended, abstract actions in an SMDP and that of the constituent primitive actions in an equivalent MDP. MAXQ makes use of a strictly hierarchical control scheme like feudal learning, but also maintains a global value function, and so somewhat resembles the HAM scheme. However, like the HAM scheme, the strength of MAXQ in tackling complex problems appears to reside in the prespecification of the hierarchical problem structure.

A different approach to hierarchical reinforcement learning has been suggested by Moore, Baird and Kaelbling (1999). A generalisation of earlier work (Kaelbling, 1993), it involves a hierarchical organisation of the state space of an MDP with a subset of the state space functioning as a set of airports ("landmarks" in Kaelbling, 1993) which other states know how to get to. Airports are arranged hierarchically, with lower level airports knowing how to get to higher level ones. Moreover, states in an airport's catchment area are assumed to be contiguous, *ie* no other states have a better value with respect to this airport as a goal. Thus for navigation worlds the airport hierarchy represents a segmentation, at different spatial scales, of the state space. This airport system is applied by Moore *et al* to the problem of learning an all-to-all value function, $V(x, y)$, where x is the current state, and y is the current goal state. Hence, as with feudal Q learning, the object is to show that the system can do better than simply learning and storing a value for every combination of x and y.

The system is model-based, *ie* involves the explicit learning of transition functions, and the use of variants of dynamic programming to calculate values. The learning issues in the model are the allocation of catchment areas to airports, and the calculation of optimal values $V(x, y)$ for states within those areas. Catchment areas for a particular airport are grown out from the closest states, but without necessarily knowing what will happen to transitions that leave the catchment area. Moore *et al* suggest a scheme which places upper and lower bounds on the values of states within a catchment area during this growing process, which are based on two separate models (best-case and worst-case) of what might happen if a transition leaves the area. Eventually, sufficient states are found that these two bounds converge to within a predefined acceptable range, and estimated optimal values are assigned to the states by averaging the best-case and worst-case values. However, it is worth pointing out two aspects of this process. First, new airport states are selected to be as far from existing airports, in terms of values, as possible. Second, the process of growing a catchment area is best suited to areas surrounded by a buffer zone of outside states where the act of leaving the area is likely to lead to a re-entry into the area. Taken together, these features imply that it is unlikely that the resulting hierarchical decomposition of the

state space will reflect underlying structure. Thus, although the airport hierarchy has more guarantees of success than feudal Q learning (which clearly fails in some cases), it is not necessarily the most effective way of constructing a structural abstraction hierarchy.

In summary, these methods, together with feudal learning, pay scant regard to environmental structure, and the fact that in general such structure has to be learned. The preliminary work described in this chapter might be useful in augmenting such methods appopriately, although it is not clear what the distinct advantages of explicitly hierarchical control are.

Chapter 7

General Conclusions

7.1 Summary of Thesis

This thesis has aimed to study navigation from a computational perspective. The first step, therefore, was an attempt to define navigation, and identify the key computational problems inherent in this definition (chapter 1). A second aim, however, has been to understand the contribution of the hippocampus to navigation. Therefore, a review was made of evidence concerning the involvement of this structure in navigational learning (chapter 2). In particular, it was noted that although the evidence for a necessary role for the hippocampus in navigational learning is strong, the nature of this role remains a controversial subject. Central to the controversy is the apparent phenomenon of place cells, that is, neurons in the hippocampus that fire in a spatially localised fashion within environments.

A suggestion made in this thesis, following similar suggestions made elsewhere (Dayan, 1991; Brown and Sharp, 1995), was that hippocampal place cells might be contributing to navigational learning by providing an appropriate representation for the application of general, reward-based learning methods. However, the theoretical sophistication of such methods has increased rapidly in recent years in a field of study somewhat removed from experimental neuroscience, that of reinforcement learning. In particular, computationally efficient, neurally plausible methods exist for learning to predict distant rewards, in ways that suggest compatibility with a place cell-like representation. A selection of these methods has therefore been reviewed (chapter 3).

A model of hippocampally dependent navigation has been presented that uses place cells as a representation of state, within a predictive learning architecture (chapter 4). The first component of the model uses temporal difference learning (Sutton, 1988) in an actor-critic scheme (*eg* Barto, Sutton and Watkins, 1990), and learns appropriate actions in a simulated RMW task. Acquisition is as fast as for rats. The second component of the model applies

temporal difference in a novel way, to learn a set of globally consistent coordinates across the watermaze environment, from local self-motion information. However, coordinate control is made to compete with conventional control, within the established mechanisms of the actor-critic, achieving appropriate switching without an explicitly separate switching mechanism. This complete model captures learning in the DMP task, in particular the gradual acquisition across days of one-trial learning.

Two purely behavioural predictions follow from the model, and were tested experimentally. First, once the putative coordinate system has been learned, simple placement of an animal at a novel goal should provide the animal with sufficient information to allow direct paths on the next trial. This was shown to be true, and in contrast to previous results (Whishaw, 1991), placement was shown to be as good as swimming for one-trial learning. The latencies of rats in the different experiments suggest that in the Whishaw study, true one-trial learning had not been achieved. A second prediction of the model follows from the weakness of a coordinate system for planning in structured environments. The model predicts that animals will be unable to learn a DMP task in the presence of barriers. In fact, a DMP barrier experiment, the H-maze, revealed one-trial learning to novel goal locations, irrespective of whether the platform was on the near or far side of a barrier. An analysis of the heading of the animal revealed that, in the first 5s of the second trial to a novel platform position, rats headed appropriately either toward the platform or around the barrier, depending on the platform's position with respect to the barrier. Furthermore, in a post-training test in which the barrier was moved between trials 1 and 2, so as to change the side which the platform was on, rats first went around the barrier or straight up to it, as would have been appropriate for the original barrier position, and then headed for the allocentric position of the platform in the pool, so that they actually bumped repeatedly into the barrier in an apparenbt attempt to swim through it. This behaviour suggested a hierarchical structure to their strategy, in attempting to circumnavigate the barrier first, and locate the platform second, and suggested that their search behaviour was not wholly dependent on the locally perceptible barrier.

Finally, several considerations (including the H-maze result decsribed above) suggested the need to consider the issue of generalisation in reinforcement learning, focusing in particular on the learning of multiple goals problems in complex environments (chapter 6). The idea was explored that optimal value functions across a subset of goals (such as might be generated for individual goals by temporal difference learning in an actor-critic scheme) might exhibit structure of the very sort which might be capitalised upon to generalise across tasks. Therefore, unsupervised learning was applied to optimal value functions (in the form of the Expectation-Maximisation algorithm; Dempster *et al*, 1977). Decompositions were found of three different types: flat, linear and hierarchical. Results show that decompositions

can be learned that correspond to the apparent structure of the world, that structure learned from just a subset of goals compares well with that learned from all possible goals, and that augmenting a standard state representation with these decompositions leads to faster acquisition using standard reinforcement learning methods.

In the following sections, some of the issues and questions raised by this work are discussed, particularly in relation to alternative theories of animal navigation. A subsequent section addresses possibilities for future work, in two parts, the first considering extensions to the reinforcement learning work described, and the second identifying possible experiments at behavioural and neurophysiological levels that are suggested by the work in this thesis.

7.2 Global Representations

As discussed in chapter 1, among behavioural neuroscientists animal navigation is usually conceived of in terms of building and using a global representation of the environment, in the form of a map. This thesis raises a number of questions about global representations, as discussed in the following sections.

No Computational Justification For A Map

It was argued in chapter 1 that a global representation such as a map of an environment presents at least two problems to its user. First, building a map requires at the very least a globally consistent coordinate system, so that information may be placed into the map in a globally meaningful way. This was described as a "global-from-local" problem. Second, reading a map presents a "local-from-global" problem, in terms of using the globally represented information to specify behaviour locally. While in restricted circumstances (such as the watermaze environment), a global bearing can be useful even at a local level, in most navigation tasks this is unlikely to be the case. In this sense, a map is a poor representation for navigation.

The problems of using a map might be less important were it to be the only way of learning about the world in order to navigate. Therefore, it is critical to note that very simple reinforcement learning methods exist for solving navigation-like problems, which also directly address the local-from-global problem, ultimately learning a local representation of the *value* of different behaviours in terms of their ultimate result. The focus of this thesis has been primarily on those aspects of reinforcement learning that have appeared problematic in relation to animal learning: speed of learning, and generalisation to similar but novel

tasks. Both problems can be addressed by assuming an appropriate local representation.

No Neural Justification For A Map

Since O'Keefe and Nadel (1978), the theory of the cognitive map has been more than merely a psychological theory of navigation, but also a neural hypothesis about the function of neurons in an area of the mammalian brain known as the hippocampus. However, from the perspective of a map, the activity of these neurons during spatial behaviour is somewhat paradoxical. Although clearly representative of spatial location within environments, these neurons apparently fail to indulge in the kind of "map"-like behaviour that might have been expected, such as looking ahead along different possible routes, or towards goal positions. Although in fact the range of navigational experiments in which place cells have been recorded is rather limited (section 7.4.2), the evidence so far collected suggests that, if place cells constitute a map, they are a map which may be only consulted to look up ones current position. Looking ahead does not appear to be allowed.

Coordinate Learning: A Global-From-Local Problem

One remaining way in which place cells might constitute a map is if the activity of place cells at different locations were associated with a representation of global coordinates, perhaps in some other neural area. However, such a representation is likely to be learned, and, as already noted, is likely to be of use only in highly restricted circumstances.

Hippocampally-dependent global coordinates can indeed be learned, at least in an environment the size of a standard watermaze, and quickly, using self-motion information and a temporal difference learning rule. Importantly, this coordinate learning is relatively independent of the particular behaviour of the animal in the watermaze, so that coordinates can be developed while nevertheless attempting to navigate to particular goals (such as in the DMP task). This exploratory freedom is not available for other global-learning schemes, such as Dayan's successor representation (Dayan, 1993).

The Computational Status Of Metric Representations

This coordinate learning raises a number of questions about the status accorded to metric representations by a number of authors. For instance, Gallistel (1990) notes that metric geometry, in which the metric relations between objects are available, *subsumes* lower level geometries, such as topology, in which only adjacency relations are available. By creating a parallel hierarchy of possible cognitive maps, Gallistel concludes that creatures as lowly

as desert ants have the richest cognitive maps available, since aspects of their navigational behaviour show evidence of metric relations (*eg* the ability of desert ants to navigate at an appropriate bearing back to a point of origin after first taking an arbitrarily windy path away from that origin). Similarly, Trullier *et al* (1997) argue that there exists a hierarchy of global spatial representations, and that a metric map must include all the information contained in a topological map. They too conclude that a demonstration of knowledge of metric relations constitutes evidence for a complete metric map, *ie* a map that also includes topological information.

The difficulty with this conceptualisation of spatial representation is that it places the emphasis on what information is "recoverable" (to use Gallistel's phrase) rather than on what information is made explicit. It has already been noted that a map in the common sense (which is usually a metric map) cannot specify in an immediate and explicit way routes other than direct bearings from start to goal. The coordinate learning model of chapter 4 makes the further point that an animal might learn a metric representation of its environment without learning anything at all about the topological structure of the environment. The original source of this metric information is self-motion information, attributed to a very large number of animals, including the desert ant. Because the model learns nothing about the topological structure of environments, it even makes the unlikely prediction that one-trial learning in the presence of barriers would not be supported.

Experimental Evidence For Global Representations

Coordinate learning can perhaps provide an explanation for a phenomenon in the behavioural literature that is commonly thought to provide evidence for cognitive mapping. Many experiments have investigated the ability of various species of animal to perform "novel short-cutting", that is, the taking of a short-cut across novel terrain because either the usual route is blocked, or much longer or costlier than the short cut. Results of this sort have been reported for hamsters (Chapuis *et al*, 1987), rats (Tolman, 1948), chimpanzees (Menzel, 1973) and, somewhat controversially, for bees (Gould, 1986). Bennett (1996) notes that the ability to make short cuts was inherent in the cognitive map concept of both Tolman (1948) and O'Keefe and Nadel (1978), but also claims that in all reported cases of short-cutting, a rather more trivial strategy would have sufficed, such as approaching a visible cue, or performing path integration. Nevertheless, it might be possible to design a true short-cutting experiment, so it is worth exploring the theoretical reasons for attempting to do so. Note then that the coordinate model might, under certain circumstances, specify direct shortcuts across an unknown terrain, for example if globally consistent coordinates had been learned throughout a U-shaped environment, and a route was attempted across the middle. Therefore, such short-cutting cannot be taken as evidence for a cognitive map, or

even for any representation of space more sophisticated than the simple coordinate model.

There is arguably something odd about the pre-occupation of cognitive map theory with behaviour in *novel* environments, or parts of an environment. Perhaps the idea of a map suggests usefulness in unfamiliar terrain – because this is when we would commonly use one. However, the deliberate focus of this thesis has been on how animals should represent and use information about spaces with which they are familiar; the very first paragraph of this thesis characterised navigation as a process of returning to important places. Consider, instead, the case of navigation across wholly unfamiliar terrain. Correct navigation (*ie* the taking of an optimal path) must be performed without any knowledge of the terrain, other than presumably the coordinate of the goal. A literally direct path is simply the result of a minimum of *prior expectations* about the nature of the terrain. However, these expectations might be wholly wrong. How does one cross the novel minefield? Does one strike out towards the distant goal, or look around for footprints?

Tolman did, however, investigate a different kind of short-cut, in which an animal finds a new shortest route (which may not be direct in the coordinate sense) through a perfectly familiar environment, which has changed in some local but significant way, perhaps in that a previously preferred route has been blocked. However, in both the "latent learning" experiment and the "insight" maze, which were essentially of this type, the experimental results fail to clarify whether or not rats have such an ability.

7.3 Local Representations

This thesis has focused on local representations for navigational learning. In particular, reinforcement learning methods are able to learn using a local representation of state (such as place cells), and local learning rules, achieving a local specification of behaviour, which is at the same time globally optimal.

The Effectiveness Of Place Cells

A key conclusion from the use of place cells in the model of chapter 4 was that their form makes them particularly suitable for the role of state space representation. As was noted in that chapter, reinforcement learning methods are typically reported to take thousands of learning trials to optimise values or actions in even simple navigation tasks, such as the example task of chapter 3. Because of their appropriately distributed representation, place cells lead to much faster learning, to the extent that actual behavioural data can be accounted for.

The effectiveness of place cells as a representation is arguably enhanced by the manner in which place fields adapt to suit the navigational environment, or task. As noted in chapter 2, place fields elongate along linear tracks, as well as remaining directional, both of which features may be interpreted as enhancing their role in learning local behaviours. Moreover, place fields do not straddle a barrier – further support for the notion of a navigational state space.

Tailoring General Learning Mechanisms With Appropriate Representations

The plastic properties of place fields suggest a more general approach to the question of generalisation in reinforcement learning. The approach in chapter 6 of this thesis, of attempting to find better representations which capture the underlying structure of the environment, may be seen as an extension of what place fields are known to do. The further issue of whether or not there might be parallel representations of the environment at different spatial scales has not been addressed directly in the literature, although relatively large place fields do occur (Wilson M, personal communication).

It has been demonstrated that from information discovered in the course of navigating to only a few goals in an environment, appropriate, hierarchical representations of state may be discovered. Further, these representations enhance learning and support a certain amount of generalisation. Within a true multiple-goals framework, they may even support one-trial learning. Importantly, however, these representations remain local, and simple in nature: they provide information about the current occupancy by the animal of the represented space. They do not constitute a map, nor do they require complex mechanisms in order to be used. The consequence of this is that the same, simple learning mechanisms may be used as were proposed for the model of RMW and DMP learning, but which have previously been ascribed to general learning of the sort revealed by conditioning experiments (*eg* Sutton and Barto, 1987). Thus the whole thrust of O'Keefe and Nadel's argument for a separation of map mechnanisms from associative learning mechanisms (*ie* locale vs. taxon processing respectively) is blunted somewhat – according to the current theoretical approach, more complex navigational behaviour can be supported using the same basic mechanisms but with better representations. A little evidence for an involvement of the striatum in some navigation tasks was noted in chapter 3, which may correspond to the involvement of general, predictive, reward-based mechanisms in hippocampally dependent navigational learning. The implication, however, is that lesion studies investigating the issue would have to be rather more sophisticated than the simple dissocation-based studies that are generally used.

7.4 Future Work

Many aspects of the work presented in this thesis are preliminary. Suggestions for future work are provided below, in two categories: extensions to the reinforcement learning work; and suggested experimental approaches.

7.4.1 Reinforcement Learning

Combinatorial Representations For Multiple Goals Tasks

One suggestion made in chapter 6 but not yet proven is that a multiplicative state space of current location and goal, in tandem with the representational learning algorithms defined, might support one-trial learning to novel goal locations quite naturally. Preliminary results, not presented in this thesis, suggest that this is indeed the case.

However, a combinatorial representation might be a rather inefficient way of representing state information. For example, the benefits of a multi-scale representation may lie in using large scale representations for behavioural decisions concerning distant goals, but the simple combinatorial scheme requires the representation and learning of the smaller scale combinations as well. Therefore it might be useful to investigate schemes for reducing the number of possible combinations, perhaps in a use-dependent manner.

An appealing neural aspect of a combinatorial representation is that the hippocampus would appear well placed to create it: the combinatorial nature of its processing of input information has been noted, as well as its involvement in storing information about the goal (*eg* the goal memory identified by Steele and Morris, 1999). Aspects of its intrinsic circuitry might also provide a means for studying the modification of these representations to emphasise useful combinations, in the manner suggested above.

Online Clustering Methods

The clustering methods described in chapter 6 were based on offline clustering of optimal value functions, using a rather global learning scheme, the expectation-maximisation algorithm. One might hope instead for an online scheme, that could learn from sub-optimal value functions, or even altogether different sources of structural information, perhaps using a more neurally plausible unsupervised learning scheme.

One simple means of learning clusters at the smallest scale would be to use coordinate control to inform about the accessibility of different areas. Preliminary results with a maximum likelihood scheme based on binary output from a coordinate controller ("success" or

"failure"; results not described in this thesis) show that clusters can be learned online and fairly rapidly, while the coordinate control specifies random choices of coordinate action. This might be one simple source of online representational learning that is also not dependent on knowing optimal values. However, yet more efficient clustering schemes might be possible.

7.4.2 Behaviour and Electrophysiology

Navigationally Demanding Place Cell Studies

Despite the hippocampus having been implicated in navigational learning through lesion, pharmacological and genetic studies, single-unit electrophysiological studies of the hippocampus in freely moving animals have generally focused on the activity of place cells in navigationally trivial tasks, many of which are even likely to be hippocampally independent, such as random exploration of open arenas, or multiple traversals of a single route or track. Partly this is because of the difficulty of achieving sufficiently high sampling of cell activity over large or diverse areas, but an additional reason is likely to be the theoretical notion of task-independent map learning developed by O'Keefe and Nadel (1978).

Many experimental results suggest that in fact task is rather important in determining the way place cells fire (*eg* Markus *et al*, 1995; Wood *et al*, 1999). Similarly, the work in this thesis suggests that some representation of the goal location may be both necessary, and mediated in some way by hippocampal neurons. For example, hippocampal neurons might show evidence for an explicit recall of goal-related place information, consonant with the delay-dependent impairment of one-trial learning in the DMP task resulting from blocked hippocampal synaptic plasticity, as found by Steele and Morris (1999). Alternatively, hippocampal neurons may develop a combinatorial representation of location and goal, and perhaps local representations of state at a hierarchy of scales. Alternatively again, truly predictive sequences of activity might be elicited, in contrast to many of the ideas presented in this thesis.

An appropriate experimental approach to the problem might combine a number of tecniques: (1) A behavioural task that developed one-trial learning to novel goal positions would be especially useful. Such a task might be developed using a lattice of tracks, so as to address the sampling problem, but also to allow recording from place cells in a single location but under multiple goal conditions. (2) Powerful, multiple single-unit recording techniques would be useful, in order to obtain sufficient numbers of cells to study. An important feature of the latest technology is that populations of cells can be held for several days, allowing cells to be recorded in many sessions, and hence, in many different condi-

tions (Dr. Matthew Wilson, Dept. of Brain and Cognitive Sciences, M.I.T., pers. comm.) Various effects might be sought: a predictive or mnemonic role for theta precession (chapter 2), clarified by varying routes just preceding or just following a particular location; location-specific planning in the form of activity of cells with place fields near the goal at times other than when the animal is at the goal, by deliberately placing goals within place fields; and a conjunctive representation of current location and goal, although the use of novel goals makes the form of such a representation were it found particularly interesting. (3) Pharmacological techniques, such as the use of an NMDA receptor antagonist, could be used to examine the dependence of any effects that were found on hippocampal synaptic plasticity. For example, if in a DMP-like task location-specific planning were found related to the current location of the goal, the natural question to ask would be whether such activity on trial 2 requires normal hippocampal synaptic plasticity on trial 1, and in a delay dependent manner, thus reminiscent of the behavioural result of Steele and Morris (1999).

Bibliography

Abbott LF, Blum KI (1995) Functional significance of long-term potentiation for sequence learning and prediction. *Cerebral Cortex* **6**:406-416.

Abbott LF, Varela JA, Sen K, Nelson SB (1997) Synaptic depression and cortical gain control. *Science* **275**:220-224.

Abraham L, Potegal M, Miller S (1983) Evidence for caudate nucleus involvement in an egocentric spatial task: return from passive transport. *Physiological Psychology* **11**(1):11-17.

Albus JS (1981) *Brains, behavior and robotics* Ch 6, pp 139-179. Byte books.

Alyan SH, Paul BM, Ellsworth E, White RD, McNaughton BL (1997) Is the hippocampus required for path integration? *Society for Neuroscience Abstracts* **23**: 504.

Amaral DG, Witter MP (1995) The hippocampal formation. In Paxinos G, ed, *The rat nervous system, 2nd edition* Academic Press.

Austin KB, White LH, Shapiro ML (1993) Short- and long-term effects of experience on hippocampal place fields. *Society for Neuroscience Abstracts* **19**:797.

Bannerman DM, Good MA, Butcher SP, Ramsay M, Morris RGM (1995) Distinct components of spatial learning revealed by prior training and NMDA receptor blockade. *Nature* **378**:182-186.

Barnes CA (1979) Memory deficits associated with senescence: a neurophysiological and behavioral study in the rat. *Journal of Comparative and Physiological Psychology* **93**:74-104.

Barto AG, Sutton RS, Anderson CW (1983) Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* **13**:834-846.

Barto AG, Sutton RS, Watkins, CJCH (1990) Learning and sequential decision making. In Gabriel M, Moore J, eds, *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge, MA, MIT Press: Bradford Books.

Bellman RE (1957) *Dynamic Programming* Princeton, New Jersey: Princeton University Press.

Bennett ATD (1996) Do animals have cognitive maps? *Journal of Experimental Biology* **199**:219-224.

Bertsekas DP (1995) *Dynamic programming and optimal control.* Belmont, MA: Athena Scientific.

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-dynamic programming.* Belmont, MA: Athena Scientific.

Biegler R (1996) *Short and Medium Range Navigation and Its Relationship To Cognitive Mapping and Associative Learning* PhD Thesis, Edinburgh University, U.K.

Biegler R, Morris RGM (1993) Landmark stability is a prerequisite for spatial but not discrimination learning. *Nature* **361**:631-633.

Bishop CM (1995) *Neural networks for pattern recognition* Oxford: Clarendon Press.

Bliss TVP, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**:31-39.

Blodgett HC (1929) The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology* **4**:113-134.

Blum KI, Abbott LF (1996) A model of spatial map formation in the hippocampus of the rat *Neural Computation* **8**:85-93.

Borovski WM (1927) Experimentelle Untersuchungen über den Lernprozess. *Zsch. vergl. Physiol.* **6**:489-529. Cited in Maier and Schneirla (1935).

Boyan JA, Moore AW (1995) Generalization in reinforcement learning: safely approximating the value function. In Cowan JD, Tesauro G, Touretzky D, eds, *Advances in Neural Information Processing Systems* **7**:369-376.

Bradtke SJ, Duff MO (1995) Reinforcement learning methods for continuous-time Markov decision processes. In Tesauro G, Touretzky DS, Leen TK, eds, *Advances in Neural Information Processing Systems* **7**.

Bridle J (1989) Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Fogelman-Soulie F, Herault J, eds, *Neuro-computing: algorithms, architectures and applications*. New York: Springer-Verlag.

Broomhead DS, Lowe D (1988) Multivariate functional interpolation and adaptive networks. *Complex Systems* **2**:321-355.

Brown EN, Frank LM, Tang D, Quirk MC, Wilson MA (1998) A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience* **18**:7411-7425.

Brown MA, Sharp PE (1995) Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. *Hippocampus* **5**:171-188.

Brown MF, Bing MN (1997) In the dark: spatial choice when access to spatial cues is restricted. *Animal, Learning & Behavior* **25**(1):21-30.

Brunel N, Trullier O (1998) Plasticity of directional place fields in a model of rodent CA3. *Hippocampus* **8**:651-665.

Buel J (1934) The linear maze. I. "Choice-point expectancy", "correctness" and the goal gradient. *Journal of Comparative Psychology* **17**:185-199.

Bunsey M, Eichenbaum H (1996) Conservation of hippocampal memory function in rats and humans. *Nature* **379**:255-257.

Burgess N, Recce M, O'Keefe J (1994) A model of hippocampal function. *Neural Networks* **7**:1065-1081.

Caramanos Z, Shapiro ML (1994) Spatial memory and N-methyl-D-aspartate receptor antagonists APV and MK-801: memory impairments depend on familiarity with the environment, drug dose and training duration. *Behavioral Neuroscience* **108**(1):30-43.

Cartwright BA, Collett TS (1983) Landmark learning in bees: experiments and models. *Journal of Comparative Physiology* **151**:521-543.

Cartwright BA, Collett TS (1987) Landmark maps for honeybees. *Biological Cybernetics* **57**:85-93.

Chapuis N, Thinus-Blanc C, Poucet B (1983) Dissociation of mechanisms involved in dogs' oriented displacements. *Quaterly Journal of Experimental Psychology* **35B**:213-219.

Chew GL, Sutherland RJ, Whishaw IQ (1989) Latent learning does not produce instantaneous transfer of place navigation: a rejoinder to Keith and McVety. *Psychobiology* **17**:207-209.

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *JASA* **83**(403):596-610.

Collett TS, Cartwright BA, Smith BA (1986) Landmark learning and visuo-spatial memories in gerbils. *Journal of Comparative Physiology A* **158**:835-851.

Colombo PJ, Davis HP, Volpe BT (1989) Allocentric spatial and tactile memory impairments in rats with dorsal caudate lesions are affected by preoperative behavioral training. *Behavioral Neuroscience* **103**:1242-1250.

Cressant A, Muller RU, Poucet B (1997) Failure of centrally placed objects to control firing fields of hippocampal place cells. *Journal of Neuroscience* **17**(7):2531-2542.

Dayan P (1991) Navigating through temporal difference. In Lippmann RP *et al*, eds, *Advances in Neural Information Processing Systems* **3**:464-470.

Dayan P (1992) The convergence of TD(λ) for general λ. *Machine Learning* **8**:341-362.

Dayan P (1993) Improving generalisation for temporal difference learning: the successor representation. *Neural Computation* **5**:613-624.

Dayan P, Abbott LF (2000) *Theoretical Neuroscience* Cambridge, MA: MIT Press, in press.

Dayan P, Hinton GE (1993) Feudal reinforcement learning. In Giles CL, Hanson SJ, Cowan JD, eds, *Advances in Neural Information Processing Systems* **5**:271-278.

Dayan P, Sejnowski TJ (1994) TD(λ) converges with probability 1. *Machine Learning* **14**:295-301.

Dayan P, Singh SP (1996) Improving policies without measuring merits. In Touretzky DS, Mozer MC, Hasselmo ME, eds, *Advances in Neural Information Processing Systems* **8**:1059-1065.

Deadwyler SA, Bunn T, Hampson RE (1996) Hippocampal ensemble activity during spatial delayed-nonmatch-to-sample performance in rats. *Journal of Neuroscience* **16**:354-372.

Dean T, Lin SH (1995) Decomposition techniques for planning in stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)* pp 1121-1127. Montreal, Canada: Morgan Kaufman.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39B**:1-38.

Devan BD, Goad EH, Petri HL (1996) Dissociation of hippocampal and striatal contributions to spatial navigation in the water maze. *Neurobiology of Learning and Memory* **66**:305-323.

Dietterich TG (1998) Hierarchical reinforcement learning with the MAXQ value function decomposition. In Shavlik J, ed, *International Conference on Machine Learning*.

Dingledine R (1983) N-methyl aspartate activates voltage-dependent calcium conductance in rat hippocampal pyramidal cells. *Journal of Physilogy (London)* **343**:385-405.

Eichenbaum H (1996) Is the rat hippocampus just for "place"? *Current Opinion in Neurobiology* **6**:187-195.

Eichenbaum H, Kuperstein M. Fagan A, Nagode J (1987) Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odour-discrimination task. *Journal of Neuroscience* **7**:716-732.

Eichenbaum H, Stewart C, Morris RGM (1990) Hippocampal representation in place learning. *Journal of Neuroscience* **10**(11):3531-3542.

Elliott MH (1929) The effect of appropriateness of reward and of complex incentives on maze performance. *University of California Publications in Psychology* **4**:91-98.

Fenton AA, Muller RU (1997) How two cues interact to conjointly control place cell firing fields. *Society for Neuroscience Abstracts* **23**:502.

Foster DJ, Morris RGM, Dayan P (1998) Hippocampal model of rat spatial abilities using temporal difference learning. In, Jordan MI, Kearns MJ, Solla SA, eds, *Advances in Neural Information Processing Systems* **10**:145-151.

Foster DJ, Morris RGM, Dayan P (1999) A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, in press.

Frey U, Morris RGM (1998) Synaptic tagging: implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences* **21**(5):181-188.

Gallistel CR (1990) *The Organization of Learning.* Cambridge, MA:MIT Press.

Georges-Francois P, Rolls ET, Robertson RG (1999) Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. *Cerebral Cortex* **9**(3):197-212.

Gerstner W, Abbott LF (1996) Learning Navigational Maps Through Potentiation and Modulation of Hippocampal Place Cells. *Journal of Computational Neuroscience* **4**:79-94.

Gluck MA, Myers CE (1996) Integrating behavioral and physiological models of hippocampal function. *Hippocampus* **6**:643-653.

Good M, Honey RC (1991) Conditioning and contextual retrieval in hippocampal rats. *Behavioral Neuroscience* **105**:499-509.

Hardy RL (1971) Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysics Research* **76**:1905-1915.

Harlow HF (1959) Learning set and error factor theory. In Koch S, ed, *Psychology: a study of a science, vol. 2*, 492-537.

Hasselmo ME, Schnell E (1994) Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. *Journal of Neuroscience* **14**:3898-3914.

Hauskrecht M (1998) Planning with temporally abstract actions. TEchnical Report CS-98-01, Department of Computer Science, Brown University, Providence, RI.

Hauskrecht M, Meuleau, Boutilier C, Kaelbling LP, Dean T (1998) Hierarchical solution of Markov-decision processes using macro-actions. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence.*

Hebb DO (1949) *The organization of behavior* New York: Wiley.

Hill AJ (1978) First occurrence of hippocampal spatial firing in a new environment. *Experimental Neurology* **62**:282-297.

Hinton GE (1987) Learning translation invariant recognition in massively parallel networks. In de Bakker JW, Nijman AJ, Treleaven PC, eds, *Proceedings PARLE Conference on Parallel Architectures and Languages Europe* pp 1-13. Berlin: Springer-Verlag.

Hinton GE, Gharamani Z (1997) Generative models for discovering sparse distributed representations *Philosophical Transactions of the Royal Society* **352B**:1177-1190.

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(3):55-67.

Hsaio HH (1929) An experimental study of the rat's "insight" within a spatial complex. *University of California Publications in Psychology* **4**:57-70.

Hull CL (1932) The goal gradient hypothesis and maze learning. *Psychological Review* **39**:25-43.

Izquierdo I, Medina JH (1998) On brain lesions, the milkman and Sigmunda. *Trends in Neuroscience* **21**(10):423-426.

Jaakkola T, Jordan MI, Singh SP (1994) On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* **6**(6).

Jarrard LE (1989) On the use of ibotenic acid to lesion selectively different components of the hippocampal formation. *Journal of Neuroscience Methods* **29**:251-259.

Jarrard LE (1993) On the role of the hippocampus in learning and memory in the rat. *Behavioral Neural Biology* **60**(1):9-26.

Jensen O, Idiart MAP, Lisman JE (1996) Physiologically realistic formation of autoassociative memory in networks with theta/gamma oscillations: role of fast NMDA channels. *Learning and Memory* **3**:243-256.

Kaelbling LP (1993) Hierarchical reinforcement learning: preliminary results. In *Proceedings of the 10th International Conference on Machine Learning* p 163. San Francisco, California: Morgan Kaufmann.

Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochstic domains. *Artificial Intelligence* **101**.

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* **4**.

Kali S, Dayan P (1998) The formation of direction independent place fields in area CA3 of the rodent hippocampus using Hebbian plasticity in a recurrent network. *Society for Neuroscience Abstracts* **24**:931.

Kearns M, Singh SP (1998) Near-optimal peformance for reinforcement learning in polynomial time. *In preparation.*

Keith JR, McVety KM (1988) Latent place learning in a novel environment and the influences of prior training in rats. *Psychobiology* **156**:286.

Kentros C, Hargreaves E, Hawkins RD, Kandel ER, Shapiro M, Muller RV (1998) Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science* **280**:2121-2126.

Knierim JJ, Kudrimoti HS, McNaughton BL (1995) Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience* **15**(3):1648-1659.

Koshland DE (1979) A model regulatory system: bacterial chemotaxis. *Physiological Review* **59**:811-862.

Lavoie AM, Mizumori SJY (1994) Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Research* **638**:157-168.

Leonhard CL, Stackman RW, Taube JS (1996) Head direction cells recorded from the lateral mammilary nuclei in rats. *Society for Neuroscience Abstracts* **22**:1873.

Levy WB (1996) A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus* **6**:579-590.

Magee JC, Johnston D (1997) A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons. *Science* **275**:209-212.

Maguire EA, Frackowiak RSJ, Frith CD (1997) Recalling routes around London: activation of the right hippocampus in taxi drivers. *Journal of Neuroscience* **17**:7103-7110.

Maier NRF (1929) Reasoning in white rats. *Comparative Psychology Monographs* **6**:93.

Maier NRF, Schneirla TC (1935) *Princples of animal psychology* New York: McGraw-Hill.

Markus EJ, Qin YL, Leonard B, Skaggs WE, McNaughton BL, Barnes CA (1995) Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience* **15**:7079-7094.

Marr D (1971) Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London, B* **262**:24-81.

Marr D (1982) *Vision* New York: Freeman.

Martin SJ, Grimwood PD, Morris RGM (2000) Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience* in press.

McClelland JL, Goddard NH (1996) Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* **6**:654-665.

McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**:419-457.

McHugh TJ, Blum KI, Tsien J, Tonegawa S, Wilson M (1996) Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. *Cell* **87**:1339-1349.

McNaughton BL, Barnes CA, Gerrard JL, Gothard K, Jung MW, Knierim JJ, Kudrimoti H, Quin Y, Skaggs WE, Suster M, Weaver KL (1996) Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *Journal of Experimental Biology* **199**:165-171.

McNaughton BL, Barnes CA, Meltzer J, Sutherland RJ (1989) Hippocampal granule cells are necessary for normal spatial learning but not for spatially-selective pyramidal cell discharge. *Experimental Brain Research* **76**:485-496.

McNaughton BL, Barnes CA, O'Keefe J (1983) The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Experimental Brain Research* **52**:41-49.

McNaughton BL, Morris RGM (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences* **10**:408-415.

McNaughton BL, Nadel L (1990) Hebb-Marr networks and the neurobiological representation of action in space. In Gluck MA, Rumelhart DE, eds, *Neuroscience and Connectionist Theory* pp 1-63. Hillsdale, New Jersey: Erlbaum.

McNaughton BL, O'Keefe J, Barnes CA (1983) The stereotrode, a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit electrodes. *Journal of Neuroscience Methods* **8**:391-397.

Mehta MR, McNaughton BL, Barnes CA, Suster MS, Weaver KL, Gerrard JL (1996) Rapid changes in the hippocampal population code during behavior: a case for hebbian learning in vivo. *Society for Neuroscience Abstract* **22**:724.15

Menzel EW (1973) Chimpanzee spatial memory organization. *Science* **182**:943-945.

Micchelli CA (1986) Interpolation of scattered data: distance matrices and conditionally positive definite functions. *COnstructive Approximation* **2**:11-22.

Miller VM, Best PJ (1980) Spatial correlates of hippocampal unit activity are altered by lesions of the fornix and entorhinal cortex. *Brain Research* **194**:311-323.

Miller WT, Glanz FH, Kraft LG (1990) CMAC: an associative neural network alternative to backpropagation. *Proceedings of the IEEE* **78**:1561-1567.

Monaghan DT, Cotman CW (1985) Distribution of N-methyl-D-aspartate-sensitive L-[3H]glutamate-binding sites in rat brain. *Journal of Neuroscience* **5**:2909-2919.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**:1936-1947.

Moore A, Atkeson CG (1995) The Parti-Game algorithm for variable resolution reinforcement learning in multi-dimensional state spaces. *Machine Learning* **21**:199-233.

Moore A, Baird L, Kaelbling LP (1998) Multi-value functions: efficient automatic action hierarchies for multiple goal MDPs. *In preparation.*

Morris RGM (1981) Spatial localisation does not require the presence of local cues. *Learning and Motivation* **12**:239-260.

Morris RGM (1983) An attempt to dissociate spatial-mapping and working-memory theories of hippocampal function. In Seifert W, ed, *The Neurobiology of the Hippocampus.* London: Academic Press.

Morris RGM (1984) Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods* **11**:47-60.

Morris RGM (1989) Synaptic plasticity and learning: selective impairment of learning in rats and blockade of long-term potentiation *in vivo* by the N-methyl-D-aspartate receptor antagonist AP5. *Journal of Neuroscience* **9**(9):3040-3057.

Morris RGM (1990) Does the hippocampus play a disproportionate role in spatial memory? *Discussions in Neuroscience* **6**:39-45.

Morris RGM, Anderson E, Lynch GS, Baudry M (1986) Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature* **319**:774-776.

Morris RGM, Garrud P, Rawlins JNP, O'Keefe J (1982) Place navigation impaired in rats with hippocampal lesions. *Nature* **297**:681-683.

Morris RGM, Schenk F, Tweedie F, Jarrard LE (1990) Ibotenate lesions of hippocampus and/or subiculum: dissociating components of allocentric spatial learning. *European Journal of Neuroscience* **2**:1016-1028.

Muller RU, Bostock E, Taube JS, Kubie JL (1994) On the directional firing properties of hippocampal place cells. *Journal of Neuroscience* **14**(12):7235-7251.

Muller RU, Kubie JL (1987) The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience* **7**:1951-1968.

Muller RU, Kubie JL, Ranck JB (1987) Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *Journal of Neuroscience* **7**:1935-1950.

Nadel L, Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal formation. *Current Opinion in Neurobiology* **7**:217-227.

Nicoll RA, Malenka RC (1995) Contrasting properties of two forms of long-term potentiation in the hippocampus. *Nature* **377**:115-118.

Nowlan SJ (1990) Maximum likelihood competitive learning. In Touretzky DS, ed, *Advances in Neural Information Processing Systems* **2**. San Mateo, CA: Morgan Kaufmann.

Nowlan SJ (1991) Soft competitive adaptation: neural network learning algorithms based on fitting statistical mixtures. Tecnical Report, CMU-CS-91-126, Carnegie Mellon University, Pittsburgh, PA.

O'Keefe J, Burgess N (1996) Geometrical determinants of the place fields of hippocampal neurons. *Nature* **381**:425-428.

O'Keefe J, Conway DH (1978) Hippocampal place units in the freely moving rat: Why they fire when they fire. *Experimental Brain Research* **31**:573-590.

O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research* **34**:171-175.

O'Keefe J, Nadel L (1978) *The hippocampus as a cognitive map.* Clarendon, London.

O'Keefe J, Recce ML (1993) Phase relationship between hippocampal place cells and the EEG theta rhythm. *Hippocampus* **3**:317-330.

Oliveira MGM, Bueno OFA, Pomarico AC, Gugliano EB (1997) Strategies used by hippocampal- and caudate-putamen-lesioned rats in a learning task. *Neurobiology of Learning and Memory* **68**:32-41.

Olton DS, Becker JT, Handelmann GE (1979) Hippocampus, space and memory. *Behavioral and Brain Sciences* **2**:313-322. See also Commentary and Response 323-366.

Olton DS, Papas BC (1979) Spatial memory and hippocampal function. *Neuropsychologia* **17**:669-682.

Olton DS, Samuelson RJ (1976) Remembrance of places past: spatial memory in rats. *Journal of Experimental Psychology: Animal Behavior Processes* **2**:97-116.

O'Mara SM, Rolls ET, Berthoz A, Kesner RP (1994) Neurons responding to whole-body motion in the primate hippocampus. *Journal of Neuroscience* **14**:6511-6523.

O'Reilly RC, McClelland JL (1994) Hippocampal conjunctive encoding, storage and recall: avoiding a tradeoff. *Hippocampus* **4**:661-682.

Packard MG, McGaugh JL (1992) Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience* **106**(3):439-446.

Panakhova E, Buresova O, Bures J (1984) Persistence of spatial memory in the Morris water maze tank task. *International Journal of Psychophysiology* **2**:5-10.

Parr R (1998) A unifying framework for temporal abstraction in stochastic processes. *Symposium on Abstraction, Reformulaton and Approximation (SARA-98).*

Parr R, Russell S (1998) Reinforcement learning with hierarchies of abstract machines. In Jordan, MI, Kearns MJ, Solla SA, eds, *Advances in Neural Information Processing Systems* **10**: 1043-1049.

Peng J, Williams RJ (1994) Incremental multi-step Q-learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp 226-232. San Francisco, CA: Morgan Kaufmann.

Poggio T, Girosi F (1990) Networks for approximation and learning. *Proceedings of the IEEE* **78**:1481-1497.

Poucet B, Thinus-Blanc C, Chapuis N (1983) Route planning in cats, in relation to the visibility of the goal. *Animal Behavior* **31**:594-599.

Powell MJD (1987) Radial basis functions for multivariable interpolation: a review. In Mason JC, Cox MG, eds, *Algorithms for approximation* pp 143-167. Oxford: Clarendon Press.

Precup D, Sutton RS (1998) Multi-time models for temporally abstract planning. In Jordan MI, Kearns MJ, Solla SA, eds, *Advances in Neural Information Processing Systems* **10**:1050-1056.

Precup D, Sutton RS, Singh SP (1998) Theoretical results on reinforcement learning with temporally abstract actions. In *Proceedings of the 10th European Conference on Machine Learning* pp 382-393. Berlin: Springer-Verlag.

Puterman ML (1994) *Markov decision processes.* New York: Wiley.

Recce ML, O'Keefe J (1989) The tetrode: a new technique for multi-unit extracellular recording. *Society for Neuroscience Abstracts* **15**:1250.

Redish AD (1997) *Beyond the cognitive map* PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Redish AD, Touretzky DS (1995) Navigating with landmarks: computing goal locations from place codes. In Ikeuchi K, Veloso M, eds, *Symbolic Visual Learning.* Oxford University Press.

Robertson RG, Rolls ET, Georges-Francois P, Panzeri S (1999) Head direction cells in the primate pre-subiculum. *Hippocampus* **9**(3):206-219.

Ross S (1983) *Introduction to stochastic dynamic programming.* New York: Academic Press.

Rudy JW, Sutherland RJ (1995) Configural association theory and the hippocampal formation: an appraisal and reconfiguration. *Hippocampus* **5**:375-389.

Rummery GA, Niranjan M (1994) On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, Cambridge.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development* **3**:210-229. Reprinted in Feigenbaum EA, Feldman J, eds, *Computers and Thought* New York: McGraw-Hill.

Saucier D, Cain DP (1995) Spatial learning without NMDA receptor dependent long-term potentiation. *Nature* **378**:186-189.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* **275**:1593-1599.

Seidenbecher T, Reymann KG, Balschun D (1997) A post-tetanic time window for the reinforcement of long-term potentiation by appetitive and aversive stimuli. *Proceedinsg of the National Academy of Sciences of the United States of America* **94**:1494-1499.

Selfridge OG (1984) Some themes and primitives in ill-defined systems. In Selfridge OG, Rissland EL and Arbib MA, eds, *Adaptive Control of Ill-Defined Systems* New York: Plenum Press.

Shallice T (1988) *From neuropsychology to mental structure* Cambridge: Cambridge University Press, pp 18-24.

Shapiro ML, Caramanos Z (1990) NMDA antagonist MK-801 impairs acquisition but not performance of spatial working and reference memory. *Psychobiology* **18**:231-243.

Shapiro ML, Tanila H, Eichenbaum H (1997) Cues that hippocampal place cells encode: dynamic and hierarchical representation of local and distal stimuli. *Hippocampus* **7**(6):624-642.

Sharp PE, Green C (1994) Spatial correlates of firing patterns of single cells in the subiculum of the freely moving rat. *Journal of Neuroscience* **14**(4):2339-2356.

Singh SP (1992a) Scaling reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth Machine Learning Conference*, pp 406-415. Morgan Kaufmann.

Singh SP (1992b) Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence.*

Singh SP, Dayan P (1998) Analytical mean squared error curves in temporal difference learning. *Machine Learning* **32**:5-40.

Singh SP, Jaakkola T, Jordan MI (1994) Learning without state estimation in partially observable environments. In *Proceedings of the Eleventh Machine Learning Conference.*

Singh SP, Jaakkola T, Jordan MI (1995) Reinforcement learning with soft state aggregation. In Tesauro G, Touretzky DS, Leen TK, eds, *Advances in Neural Information Processing Systems* **7**.

Singh SP, Sutton RS (1996) Reinforcement learning with replacing eligibility traces. *Machine Learning* **22**:123-158.

Skaggs WE, McNaughton BL, Wilson MA, Barnes CA (1996) Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**:149-172.

Speakman A, O'Keefe J (1991) Hippocampal complex spike cells do not change their place fields if the goal is moved within a cue controlled environment. *European Journal of Neuroscience* **2**:544-555.

Specht DF (1991) A general regression neural network. *IEEE Transactions on Neural Networks* **2**:568-576.

Spooner RIW, Thomson A, Hall J, Morris RGM, Salter SH (1994) The atlantis platform: a new design and further development of Buresova's on demand platform for the water maze. *Learning and Memory* **1**:203-211.

Steele RJ, Morris RGM (1999) Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus* **9**:118-136.

Stewart CA, Morris RGM (1992) The watermaze. In Sahgal A, ed, *Behavioural neuroscience: a practical approach* Oxford: IRL Press Ch. 9: 107-122.

Sutherland RJ, Arnold KA, Rodriguez AR (1987a) Anterograde and retrograde effects on place memory after limbic or diencephalic damage. *Society for Neuroscience Abstracts* **13**:1066.

Sutherland RJ, Dyck, RH (1984) Place navigation by rats in a swimming pool. *Canadian Journal of Psychology* **38**:322-347.

Sutherland RJ, Hoesing JM (1993) Posterior cingulate cortex and spatial memory: A microlimnology analysis. In Vogt BA, Gabriel M, eds, *Neurobiology of cingulate cortex and limbic thalamus: a comprehensive handbook* Boston: Birkbauer. pp 461-477.

Sutherland RJ, Rodriguez AJ (1990) The role of the fornix/fimbria and some related subcortical structures in place learning and memory. *Behavioral and Brain Research* **32**:265-277.

Sutherland RJ, Rudy J (1989) Configural association theory: the role of the hippocampal formation in learning, memory and amnesia. *Psychobiology* **17**:129-144.

Sutherland RJ, Whishaw IQ, Kolb B (1983) A behavioural analysis of spatial localisation following electrolytic, kainate or colchicine induced damage to the hippocampal formation in the rat. *Behavioral Brain Research* **7**:133-153.

Sutherland RJ, Whishaw IQ, Regehr JC (1982) Cholinergic receptor blockade impairs spatial localization by use of distal cues in the rat. *Journal of Comparative and Physiological Psychology* **96**(4):563-573.

Sutton RS (1984) *Temporal Credit Assignment In Reinforcement Learning.* PhD Thesis, Dept. of Computer and Informaton Systems, University of Massachusetts, Amherst MA.

Sutton RS (1988) Learning to predict by the methods of temporal difference learning. *Machine Learning* **3**:9-44.

Sutton RS (1995) TD models: modeling the wold at a mixture of time scales. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML-95)* pp 531-539. San Mateo, California: Morgan Kaufmann.

Sutton RS (1996) Generalization in reinforcement learning: successful examples using sparse coarse coding. In Touretzky DS, Mozer MC, Hasselmo ME, eds, *Advances in Neural Information Processing Systems* **8**:1038-1044.

Sutton RS (1999) Open theoretical questions in reinforcement learning. `http://www.cs.umass.edu/~rich/`.

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review* **88**:135-170.

Sutton RS, Barto AG (1987) A temporal-difference model of classical conditioning. *GTE Laboratories Technical Report* TR87-509.2.

Sutton RS, McAllester D, Singh S, Mansour Y (1999) Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems* **12**. In press.

Sutton RS, Pinette B (1985) The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society.*

Sutton RS, Precup D, Singh SP (1998) Between MDPs and SMDPs: Learning, planning and representing knowledge at multiple temporal scales. Technical Report 98-74, Dept. of Computer Science, University of Massachusetts, Amherst, MA.

Tanila H, Shapiro ML, Eichenbaum H (1997) Discordance of spatial representation in ensembles of hippocampal place cells. *Hippocampus* **7**(6):613-623.

Taube JS (1995) Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience* **15**:70-86.

Teyler TJ, DiScenna P (1986) The hippocampal memory indexing theory. *Behavioral Neuroscience* **100**:147-154.

Thrun S, Schwartz A (1995) Finding structure in reinforcement learning. In Tesauro G, Touretzky D, Leen T, eds, *Advances in Neural Information Processing Systems* **7**.

Tolman EC (1949) *Purposive behavior in animals and men* Berkeley: University of California Press.

Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* **55**:189-208.

Tolman EC (1951) *Collected papers in psychology.* Berkeley: University of California Press.

Tolman EC, Honzik CH (1930a) "Insight" in rats. *University of California Publications in Psychology* **4**:215-232.

Tolman EC, Honzik CH (1930b) Degrees of hunger, reward and non-reward, and maze learning in rats. *University of California Publications in Pschology* **4**:241-256.

Trullier O, Wiener SI, Berthoz A, Meyer J (1997) Biologically-based artificial navigation systems - review and prospects. *Progress in Neurobiology* **51**(5):483-544.

Tsien JZ, Chen DF, Gerber D, Tom C, Mercer EH, Anderson DJ, Mayford M, Kandel ER, Tonegawa S (1996a) Subregion- and cell type-restricted gene knockout in mouse brain. *Cell* **87**:1317-1326.

Tsien JZ, Huerta J, Tonegawa S (1996b) The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. *Cell* **87**:1327-1338.

Tsitsiklis JN, Van Roy B (1996) Feature-based methods for large scale dynamic programming. *Machine Learning* **22**:59-94.

Tsitsiklis JN, Van Roy B (1997) An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control* **42**(5):674-690.

Wan HS, Touretzky DS, Redish AD (1993) Towards a computational theory of rat navigation. In: *Proceedings of the 1993 Connectionist Models Summer School.* Hillsdale, NJ: Lawrence Erlbaum, 11-19.

Watkins CJCH (1989) *Learning from delayed rewards.* PhD thesis, King's College, Cambridge, U.K.

Watkins CJCH, Dayan P (1992) Q-Learning. *Machine Learning* **8**:279-292.

Weisend MP, Astur RS, Sutherland RJ (1996) The specificity and temporal characteristics of retrograde amnesia after hippocampal lesions. *Society for Neuroscience Abstracts* **22**:1118.

Whishaw IQ (1985a) Formation of a place learning set in the rat: a new paradigm for neurobehavioral studies. *Physiology and Behavior* **35**:139-145.

Whishaw IQ (1985b) Cholinergic receptor blockade in the rat impairs locale but not taxon strategies for place navigation in a swimming pool. *Behavioral Neuroscience* **99**(5):979-1005.

Whishaw IQ (1991) Latent learning in a swimming pool place task by rats: evidence for the use of associative and not cognitive mapping processes. *Quarterly Journal of Experimental Psychology* **43B**(1):83-103.

Whishaw IQ, Jarrard LE (1996) Evidence for extrahippocampal involvement in place learning

and hippocampal involvement in path integration. *Hippocampus* **6**:513-524.

Whishaw IQ, Mittleman G, Bunch TS, Dunnett SB (1987) Impairments in the acquisition, retention and selection of spatial navigation strategies after medial caudate-putamen lesions in rats. *Behavioural Brain Research* **24**:125-138.

Widrow B, Hoff ME (1960) Adaptive switching circuits. *1960 WESCON Convention Record, Part IV* pp 96-104.

Wiener SI (1993) Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *Journal of Neuroscience* **13**:3802-3817.

Wiener SI (1996) Spatial, behavioral and sensory correlates of hippocampal CA1 complex spike cell activity: implications for information processing functions. *Progress in Neurobiology* **49**:335-361.

Wiener SI, Paul CA, Eichenbaum H (1989) Spatial and behavioral correlates of hippocampal neuronal activity. *Journal of Neuroscience* **9**:2737-2763.

Wilkie DM, Palfrey R (1987) A computer simulation model of rats' place navigation in the Morris water maze. *Behavior Research Methods, Instruments and Computers* **19**(4):400-403.

Wilson ED (1971) *The Insect Societies* Cambridge: Belknap.

Wilson MA, McNaughton BL (1993) Dynamics of the hippocampal ensemble code for space. *Science* **261**:1055-1058.

Witten IH (1977) An adaptive optimal controller for discrete-time Markov environments. *Information and Control* **34**:286-295.

Wood ER, Dudchenko PA, Eichenbaum H (1999) The global record of memory in hippocampal neuronal activity. *Nature* **397**:613-616.

Worden R (1992) Navigation by fragment fitting: a theory of hippocampal function. *Hippocampus* **2**(2):165-188.

Zeldin RK, Olton DS (1986) Rats acquire spatial learning sets. *Journal of Experimental Psychology: Animal Behavior Processes* **12**:412-419.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience* **16**(6):2112-2126.

Zipser D (1986) Biologically plausible models of place recognition and goal location. In Rumelhart DE, McClelland JL, ed.s *Parallel Distributed Processing, Volume 1.* MIT Press, 1986; Chp 23.